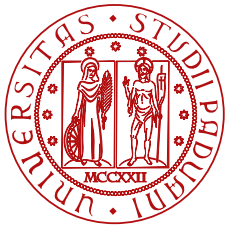


Università degli Studi di Padova

Dipartimento di Ingegneria dell'Informazione

Master Thesis in ICT for Internet and Multimedia

Study and Implementation of Multimodal Generative AI Systems for Emotion-Conditioned Music Generation

Master Candidate

Ismail Deha Kose

Student ID 2072544

Supervisor

Prof. Antonio Rodà

University of Padova

Co-supervisor

Dott. Matteo Spanio

University of Padova

Academic Year
2024/2025

To my family,
For being my safe haven throughout this long journey, for your unwavering support and your unconditional faith in me.

To my friends,
For your exceptional friendships that made my Masters studies more manageable and more meaningful, never withholding your supporting, understanding and joy even in the most difficult moments.

Your presence made all the difference.

Abstract

This thesis presents the development of a multimodal generative artificial intelligence system capable of producing emotionally consistent content across visual and auditory data. Its main goal was to build a model that can analyze the emotional content perceived (specifically Valence and Arousal) from food images and then synthesize the original audio reflecting that emotional state. To accomplish this goal, a model was developed based on a Variational Autoencoder (VAE) architecture, adopting a "token-to-token" paradigm. This research utilized the FoodPics Extended 2022 and DEAM datasets, both of which include the necessary emotional annotations.

The research systematically compared two primary architectural strategies. The first involved models using modality-specific tokenizers (ViT for images, EnCodec for audio), which required complex dimensional adaptation techniques. The second approach, in contrast, converted audio signals into mel-spectrograms, allowing a unified Vision Transformer (ViT) to process both visual and auditory data.

Experimental results demonstrated that the unified tokenizer architecture was significantly superior in learning cross-modal representations and reducing architectural complexity. The most significant breakthrough among the six model versions was achieved in Version 3.2 with the introduction of a novel "CLS-informed" generative mechanism. This innovative architecture processes the global semantic information from the ViT's [CLS] token separately from the patch tokens, then uses this global context to conditionally guide the reconstruction of local details.

This approach yielded a significant improvement, reducing the reconstruction loss than 0.025, an enhancement of over 90% compared to earlier versions. Furthermore, the model proved its ability to accurately predict Valence-Arousal values from the latent space and to meaningfully organize this space according to both semantic content and emotional values.

In conclusion, this thesis successfully demonstrates the feasibility of developing emotion-aware cross-modal generative systems. It shows the effectiveness of the unified representation learning approach, establishing a solid foundation for future multimodal AI systems.

Contents

List of Figures

List of Tables

List of Acronyms

AI Artificial Intelligence

GenAI Generative Artificial Intelligence

GAN Generative Adversarial Network

VAE Variational Autoencoder

VA Valence-Arousal

ViT Vision Transformer

DEAM Database for Emotional Analysis in Music

1 Introduction

The rapid growth of artificial intelligence has changed how we think about how machines can see, understand, and make multimedia content. One of the most exciting developments in recent years has been the emergence of multimodal generative systems. These systems can bridge the gap between different sensory modalities and establish meaningful connections between visual, auditory, and textual information. They represent an important step in computational creativity and offer new possibilities for human-computer interaction and content generation, which were previously theoretical.

Traditional artificial intelligence approaches have generally focused on unimodal systems that process a single type of input; for example, language models that only process text or classification systems that only process images. Nevertheless, human perception is fundamentally multimodal. To understand our environment completely and contextually, we excellently integrate information from multiple senses. This understanding has led researchers to develop increasingly complex multimodal AI systems that can replicate this integrated information processing approach [1].

1.1 Generative Artificial Intelligence

Generative Artificial Intelligence (GenAI) represents a new perspective from traditional discriminative models that classify or predict existing patterns to models that can create entirely new content. Generative models, on the other hand, learn the underlying probability distributions of the training data. conven-

tional AI systems focus on recognizing patterns in data, but generative models learn the underlying probability distributions of the data. They use this information to make new samples that are similar to the original dataset [2].

The field has seen major breakthroughs in recent years. Models now demonstrate notable capabilities in generating high-quality images, realistic speech, coherent text, and even complex multimedia content. Sophisticated architectures have driven these advances. These include Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs). More recently, diffusion models and transformer-based approaches have emerged [3].

Recent developments have shown that generative models can perform representation learning and generative modeling within latent spaces. This provides more enhanced cross-modal understanding and generation capabilities. This evolution has been particularly notable in the development of unified frameworks capable of processing multiple modalities simultaneously. This opens up new possibilities for creative applications in various fields.

The success of GenAI has significant implications for countless fields, ranging from entertainment and media production to scientific research and education. In this research context, generative models provide the foundation for creating systems that can understand emotional content in one modality and generate corresponding expressions in another. This represents a significant step towards more intuitive and emotionally aware AI systems.

1.2 Multimodal Systems

Multimodal AI systems represent a natural evolution more human-like artificial intelligence. They process and integrate information from multiple input modalities simultaneously. These systems recognize that human cognition completely operates across sensory boundaries. We combine visual, auditory, textual, and other forms of information. This creates comprehensive understanding of complex situations. [4].

The development of multimodal systems shows that the fundamental limitations of unimodal approaches. Single-modality systems often struggle to capture interconnectedness of real-world data. These systems can learn joint representations across modalities to get a better and more useful understanding. This approach outperforms systems that process one modality at a time. [5].

Developments in multimodal learning have demonstrated significant capabilities across various applications. Vision-language models can understand and reason about images by using textual information. However, audio-visual systems can learn meaningful correspondences between sound and visual content [6]. Transformer architecture's breakthrough has been supported by these developments. They can effectively process sequences of tokens from different modalities within unified frameworks.

Multimodal VAEs have emerged as particularly powerful tools for learning joint representations across modalities. These models can effectively combine information from different sources in their latent space. This unlocks analysis and generation tasks across modalities. The multimodal representations learned can be used for transfer learning in downstream tasks. Thus, it provides frameworks for performing unconditional and conditional generation beyond different modalities.

A fundamental design choice in any multimodal system is its fusion architecture, which dictates how and when information from different data streams is integrated. These architectures are generally categorized based on the stage at which fusion occurs. Early fusion (or data-level fusion) strategies merge raw or minimally processed data at the input layer, creating a single representation that is then fed into a unified network. In contrast, intermediate fusion (or feature-level fusion) involves processing each modality through separate pathways to extract high-level features, which are then merged in the hidden layers of the model. Finally, late fusion (or decision-level fusion) strategies process each modality independently to produce separate outputs or decisions, which are then combined at the very end. The choice of fusion strategy profoundly impacts a model's complexity, performance, and its ability to learn intricate cross-modal correlations [5].

The use of multiple modalities also presents new challenges. These include the need for temporal alignment, the need to deal with missing modalities, and meaningful correspondences between different types of data. Research has made significant progress in dealing with these challenges by developing new model architectures and training strategies [7].

1.3 Advances in Image and Sound Generation

The past decade has witnessed major developments in both image and sound generation. That was establishing the foundation for enhanced cross-modal applications. In computer vision, models have demonstrated impressive capabilities in creating photorealistic images and manipulating existing images with high accuracy [8].

The introduction of Vision Transformers (ViTs) has revolutionized image processing by applying transformer architectures to visual content. ViTs treat images as sequences of patches. This enables more sophisticated understanding of complex visual patterns and relationships compared to traditional convolutional approaches [9]. Tasks requiring detailed understanding of visual content and its semantic meaning have found this advancement particularly valuable.

In the audio domain, significant breakthroughs have been achieved in both analysis and synthesis. WaveNet and other deep learning models have shown that they can make high-quality speech and music directly from raw audio waveforms [10]. Neural audio codecs like EnCodec have enabled efficient tokenization of audio in recent years. This makes it possible to represent audio as discrete token sequences [11].

Transformer-based audio models have further enhanced the field's capabilities. Models like VATT (Video-Audio-Text Transformer) have shown that convolution-free transformer architectures can learn multimodal representations from raw signals. In many downstream tasks, these models operate superior to traditional ConvNet-based methods [12]. These advances have established the technical foundation necessary for sophisticated cross-modal generation systems.

Audio generation has also explored emotion-conditioned synthesis in recent work. Models learn to generate audio content that reflects specific emotional states in this approach. This research has demonstrated that neural networks can capture and manipulate emotional characteristics in generated audio. This provides important insights for developing emotionally-aware multimodal systems [13].

1.4 Emotion Recognition and the Valence-Arousal Model

An essential aspect in developing emotionally-aware multimedia systems is the correctly representation and processing of emotions. The dimensional approach to emotion, particularly the Valence-Arousal (VA) model, has become a robust framework for comprehending and modeling emotional states. The valence dimension represents the pleasantness (positive valence) or unpleasantness (negative valence) of emotion. Arousal measures its intensity, ranging from active/exciting (high arousal) to passive/dull (low arousal) [14].

Figure 1.1: The Valence-Arousal emotion space showing different emotional states positioned according to their valence (pleasant-unpleasant, x dimension) and arousal (high-low activation, y dimension) dimensions.

This dimensional approach has proven itself with valuable in computational applications. It provides a continuous space where all emotions can represent. The VA model captures the full range of emotional experiences. However, discrete categorical models may miss small emotional differences. The VA model has been effectively utilized across multiple domains, from speech emotion recognition to visual affect analysis, demonstrating its versatility and effectiveness [15].

Speech emotion recognition has shown notable progress in recent advances. Transformer-based models now achieve high precision for both valence and arousal prediction. This progress has been particularly significant in closing the "valence gap". This gap has historically been a major challenge in speech emotion recognition. In this field, arousal prediction consistently outperformed valence prediction [16].

The application of VA modeling extends beyond speech to visual content. Researchers have developed enhanced methods for extracting emotional information from images. Visual content can reliably elicit specific emotional responses, as studies have shown. These responses can be quantified using the VA framework. This makes it possible to develop systems that understand emotional content in visual media [17].

In the context of food imagery, the VA model has proven particularly relevant due to the strong emotional associations people have with food. Research has shown that food images can produce consistent and measurable emotional responses. These responses correlate well with factors such as palatability, desire to eat, and overall affective response [18].

1.5 The Intersection of Food Imagery and Emotional Response

Food imagery represents a particularly domain for emotional analysis due to its deep connection with human psychological and physiological responses. Visual food cues have been demonstrated to guide eating behavior by eliciting food craving and influencing food choice. Their relevance in human information processing has been shown across psychology, medicine, and neuroscience [19].

The Food-Pics dataset has emerged as a crucial resource for systematic research in this area. This comprehensive database comprises 568 food images and 315 non-food images. All images include detailed metadata. It includes normative ratings from 1,988 individuals with large variance in age and weight. The dataset provides ratings for valence, arousal, palatability, desire to eat, recognizably, and visual complexity. This makes it an ideal selection for developing emotion-aware systems [18].

Standardized food images can consistently elicit specific emotional reactions. Factors such as cultural background, individual preferences, and physiologi-

cal state influence the strength and nature of these responses. These findings provide strong empirical support for developing multimodal systems that can transfer emotional content from food images to other modalities.

The emotional richness of food imagery extends beyond simple preference judgments. It encompasses complex associations with memory, culture, and personal experience. Food images provide an ideal testbed for developing sophisticated emotion recognition and transfer systems due to this complexity. They provide a natural domain where emotional content is both readily apparent and deeply meaningful.

1.6 Tokenization Approaches for Multimodal Systems

A fundamental challenge in developing multimodal AI systems lies in creating unified representations that can effectively handle different types of data. Tokenization, the process of converting continuous data into discrete tokens, has emerged as a key solution. Models can work with a variety of data types in a single framework by this method. It allows various type data to be represented as sequences of discrete tokens that can be processed by unified architectures.

Various tokenization approaches have been developed for different modalities. For visual modality, traditional methods often rely on patch based representations or convolutional features. However, Vision Transformers (ViTs) have demonstrated superior performance by treating images as sequences of patches. Complex visual understanding tasks have found this approach more effective than conventional methods [9].

For audio content, tokenization methods range from spectral representations to more advanced neural codecs. The choice of tokenization method significantly impacts the model's ability to capture and reconstruct audio content. It must maintain important characteristics such as temporal structure and frequency information. Neural audio codecs like EnCodec have shown particular promise in creating discrete representations that preserve audio quality while enabling efficient processing [11].

The development of token to token models represents a significant advance in multimodal AI. These systems can learn direct mappings between tokenized representations of different modalities. This approach emerges unique challenges.

The model must understand the individual characteristics of each modality. It must also understand the complex relationships between them. Traditional approaches might convert tokens back to its own type for cross-modal processing or might use one downstream framework for each modality. Token to token models maintain the discrete nature throughout the entire process. This enables more direct and potentially more efficient cross-modal generation.

Similar systems for emotion-conditioned cross-modal generation between food images and audio content do not currently exist. This represents the novelty of this approach. This presents both an opportunity and a challenge, as it requires developing new methodologies without established benchmarks or reference implementations.

1.7 Research Gap and Motivation

Significant progress has been recorded in multimodal AI systems. However, there are still areas that need to be remarkable in systems. It is an example to create a relevant audio content based on visual inputs. Existing research has explored various forms of cross -modal production. The special difficulty of creating sound sequences reflecting the emotional content perceived from food images has not been adequately addressed.

Current multimodal systems often focus on direct content translation rather than capturing and transferring the emotional essence of content across modalities. Such as, describing images with text or generating images from text descriptions. Emotional content transfer requires a understanding of how emotions are expressed differently across modalities. It also requires understanding how these expressions can be meaningfully connected. This represents a significant limitation [20].

The lack of systems capable of generating emotionally compatible audio from food images and represents both a technical challenge and a missed opportunity. Applications could be for computational creativity and making content that is unique to each user. Food images, with their emotional associations and cultural significance, provide an ideal domain. They offer opportunities for exploring these challenges.

Existing cross-modal generation systems operate on high-level semantic correspondences rather than emotional correspondences. This limits their ability to create content that resonates emotionally with users. It misses a crucial aspect

of human multimedia experience. Developing cross-modal generation systems that are aware of emotions is a important step toward making AI systems that are more like people and easier to use.

1.8 Potential Impact and Applications

This research has significant implications for multiple domains within computational creativity and affective computing. In music therapy, the system could make it possible to make personalized audio content based on how people feel about visual stimuli.

For personalized media recommendation systems, the technology could improve content delivery by making audio accompaniments that match the emotional tone of visual media. This would make user experiences more immersive and emotionally consistent. In interactive art installations, the system could enable real-time generation of soundscapes that respond to the emotional content of visual elements. This would create dynamic and engaging artistic experiences.

The advertising business could use the ability to make audio content that goes along with the emotional impact of visual ads. This would ensure consistent emotional messaging across modalities. Similarly, the film and gaming industries could use such systems to automatically generate appropriate audio atmospheres that match visual content. This would streamline the content creation process while maintaining emotional coherence.

This technology could also be useful in schools, especially in language learning, where emotional context is very important for understanding and remembering what you learn. These kinds of systems could make learning more fun by making audio content that matches the emotional tone of visual materials. They could also make learning more effective.

2

Emotion Based Dataset

2.1 Dataset

This section describes the data sets used in the study, their characteristics, preprocessing steps, and preparation for the multimodal learning task. Two main data sets were used: DEAM (Database of Emotion Analysis Using Music) for audio data and FoodPics Extended 2022 for visual data. Both datasets include valence-arousal (VA) emotion labels required for the emotion-conditional generation task. The data processing approach was developed in multiple versions to optimize performance and improve emotion representation.

2.1.1 Audio Dataset: DEAM

The DEAM (Database for Emotion Analysis using Music) dataset was selected for providing audio content with continuous emotion annotations. The original DEAM dataset contains 1,802 music tracks with corresponding valence and arousal values obtained through human annotation studies. This dataset provides both dynamic (time-varying) and static (song-level) emotion annotations, making it suitable for emotion analysis.

Audio Processing Evolution

The audio processing approach underwent significant changes across different model versions to address architectural requirements and improve performance:

Versions 1-2 (46-second audio approach): In the initial versions, the complete audio samples were used with a duration analysis to determine the optimal length. After systematic analysis, 46 seconds was identified as the longest duration that could be applied to most samples without requiring truncation. This decision was important for maintaining emotion annotation validity, since VA values were calculated based on complete audio content. The filtering process resulted in 1,744 samples of exactly 46 seconds each, converted to mono format at 44.1 kHz sampling rate.

Version 3 (5-second segmentation approach): The audio processing strategy was completely revised to better capture the temporal emotion dynamics. This enabled more detailed emotion modeling through segmenting longer audio files into shorter chunks, which allowed for more precise emotion annotation and better training data diversity.

Audio Segmentation and Processing (Version 3)

The Version 3 audio processing pipeline implemented a comprehensive approach to temporal emotion modeling:

Audio Segmentation Process: All 1,802 original audio samples were segmented into 5 second chunks starting from the 15th seconds. This starting point was chosen because of the dynamic annotations begin at 15th seconds. The beginning of audio files might contain distortions or issues which may not correctly represent the emotional values of the music. The segmentation process created 13,023 total chunks.

After quality assessment, 56 chunks that were shorter than 5 seconds were identified and removed to ensure consistency. Through standardization and padding processes, a final dataset of 12,974 audio chunks was obtained, with each chunk exactly 5.0 seconds in duration. This approach significantly increased the training data amount. Each original audio sample contributed multiple segments, which provided better coverage of the emotion space.

Mel-Spectrogram Generation: Audio clips were converted into mel-spectrograms using the Riffusion model, which converts audio signals into visual representations that can be processed by Vision Transformers (Riffusion). The technical specifications used are as follows:

- **Sample Rate:** 44,100 Hz
- **Frequency Parameters:** 512 frequency bins covering the full audio spectrum

- **Time Parameters:** 10ms step size with 100ms window duration
- **Output Resolution:** 512x512 pixels, then resized to 224x224 for ViT compatibility

This configuration preserves important acoustic characteristics. It creates visual representations suitable for transformer-based processing.

Audio Emotion Annotations

The DEAM dataset provides emotion annotations in two different formats:

Static Annotations (Versions 1-2): Song-level VA values representing the overall emotional characteristics of each complete audio sample. These annotations provide a single valence and arousal value for each entire song, which was used in the earlier versions for consistency with the image dataset format.

Dynamic Annotations (Version 3): VA values per second throughout each audio sample at 500ms intervals starting from the 15th second. These annotations capture how emotions change over time within music tracks and were used for track-level emotion assignment in Version 3. The corresponding dynamic VA values were extracted and averaged for each 5 second chunk. It was made to provide stable emotion representations while preserving temporal detail.

All VA values were normalized using min-max normalization to the range [-1, 1] to ensure consistent representation across both audio and visual modalities.

Figure 2.1 shows the distribution of valence and arousal values across the audio dataset, illustrating the coverage of the complete emotion space with particular concentration in the excited and content regions, which is typical for music content.

2.1.2 Visual Dataset: FoodPics Extended 2022

The FoodPics Extended 2022 dataset was chosen for its comprehensive collection of food images. Additionally, it was detailed emotion annotations across different demographic groups. This dataset provides valence and arousal ratings from multiple participant groups. It is enabling robust emotion modeling and reducing bias from individual demographic differences.

Image Selection and Filtering

The original FoodPics dataset contained 1,211 images. A systematic filtering process was applied to ensure data quality and maintain focus on the research

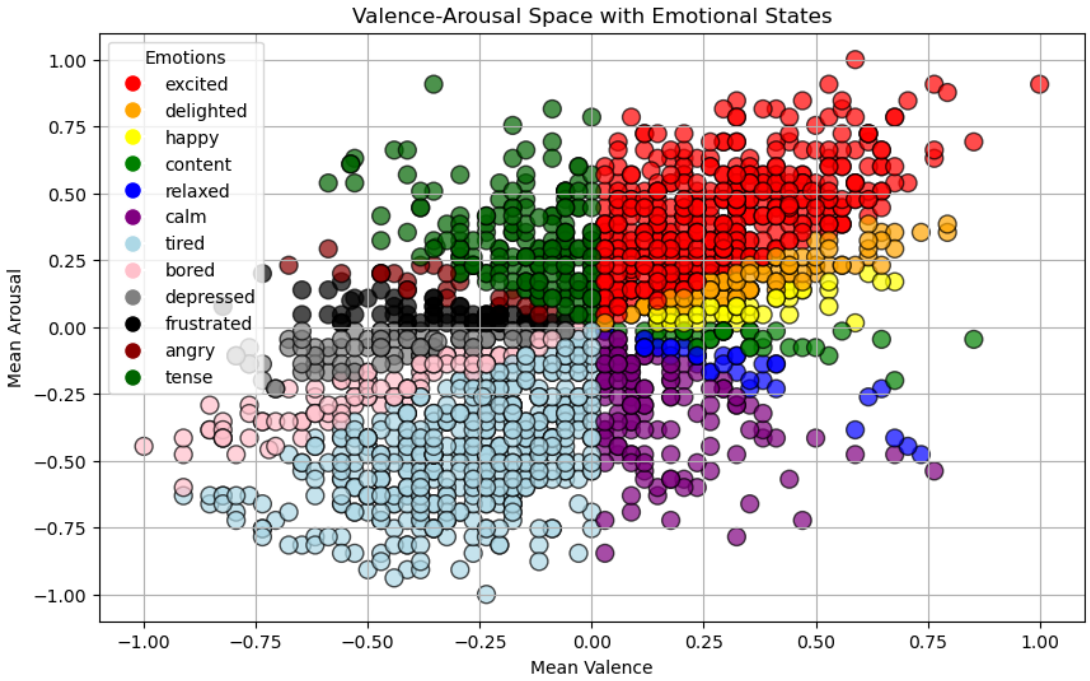


Figure 2.1: DEAM Dataset VA Distribution

domain:

1. **Category Selection:** Only food-related images were kept, while non-food items such as kitchen utensils, packaging materials, and other objects were removed to maintain focus on the core research area.
2. **Annotation Completeness:** Images lacking complete VA annotations were removed to ensure consistent emotion labeling in the dataset.
3. **Quality Control:** Images with missing data, corruption, or processing errors were excluded through systematic validation.

This filtering process ended as the removal of 145 images. It was yielding a final dataset of 1,066 high-quality food images with complete emotion annotations.

Image Emotion Annotations

The FoodPics dataset provides VA annotations from four demographic groups: Omnivore Male, Omnivore Female, Vegetarian Male, and Vegetarian Female. A unified approach was adopted rather than using demographic-specific annotations. Mean of valence and arousal values were calculated in all available demographic groups for each image. This strategy was implemented as follows:

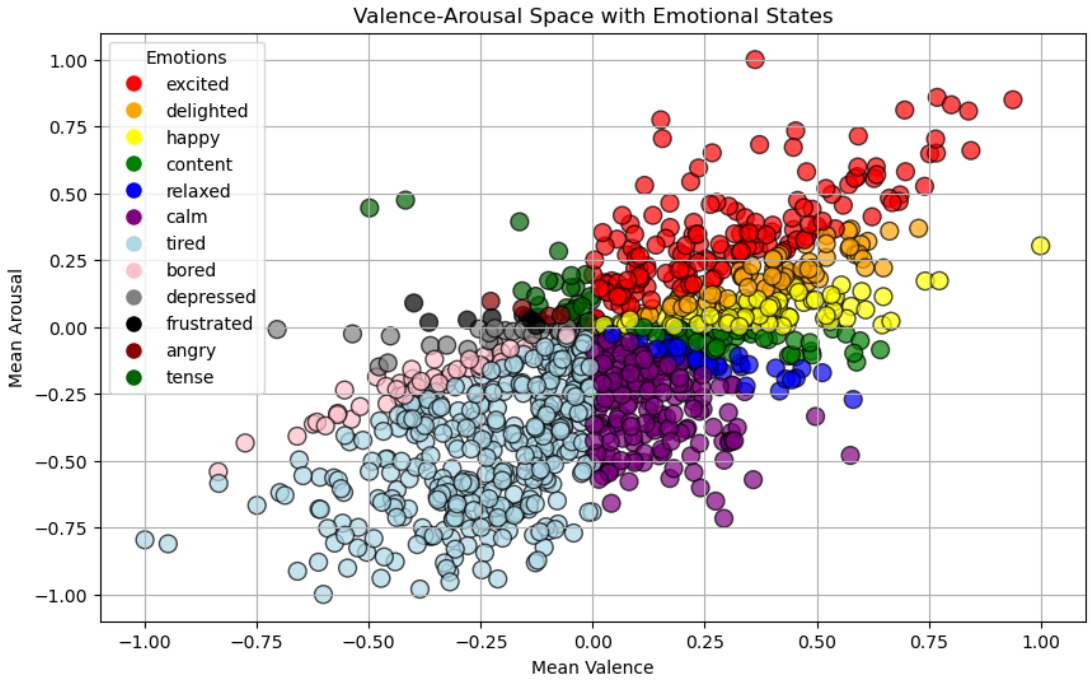


Figure 2.2: FoodPics Extended 2022 Dataset VA Distribution

```
valence_columns = ['Valence_omnivore_Male', 'Valence_omnivore_Female'
    , 'Valence_Vegetarian_Male', 'Valence_Vegetarian_Female']
Mean_Valence = mean(valence_columns)

arousal_columns = ['Arousal_omnivore_Male', 'Arousal_omnivore_Female'
    , 'Arousal_Vegetarian_Male', 'Arousal_Vegetarian_Female']
Mean_Arousal = mean(arousal_columns)
```

This approach reduces individual demographic bias. It provides more generalizable emotion labels suitable for the multimodal learning task. The emotion values were then normalized as the [-1, 1] range by using min-max normalization to ensure consistency with the audio dataset.

Figure 2.2 displays the valence-arousal distribution for the image dataset. It showed a more concentrated distribution compared to audio, with significant representation in positive valence regions, which aligns with the generally attractive nature of food images.

Image Preprocessing Pipeline

Image preprocessing was designed to meet Vision Transformer architecture requirements while preserving important visual characteristics:

1. **Resolution Standardization:** Original images (typically varying in size) were resized to 224x224 pixels using high-quality Lanczos resampling. This standard size is required for ViT processing to maintain visual details important for emotion recognition.
2. **Format Conversion:** All images were converted to RGB format without additional color normalization or any augmentation techniques. Natural color characteristics were preserved because they are important for food emotion perception.
3. **Quality Validation:**Each image has undergone validation for proper format, successful resizing, and absence of corruption before inclusion in the final dataset.

Emotion-Based Categorization and Analysis

A comprehensive emotion categorization system was implemented to the dataset. It was aimed to analyze the emotional distribution and understand the emotional patterns associated with different food types:

Emotion Mapping Algorithm: Images were categorized into emotion classes by using a slope-based approach that based on both valence and arousal dimensions. The algorithm first calculates the slope as the ratio between arousal and valence scores:

$$\text{slope} = \frac{\text{arousal\_score}}{\text{valence\_score}} \tag{2.1}$$

The classification then follows a two-step decision process based on valence polarity and slope thresholds:

For Positive Valence (valence_score $>$ 0):

- excited: slope > 0.66
- delighted: $0.33 <$ slope ≤ 0.66
- happy: $0 <$ slope ≤ 0.33
- content: $-0.33 <$ slope ≤ 0
- relaxed: $-0.66 <$ slope ≤ -0.33
- calm: slope ≤ -0.66

For Negative Valence (valence_score $\leq$ 0):

- tired: slope > 0.66

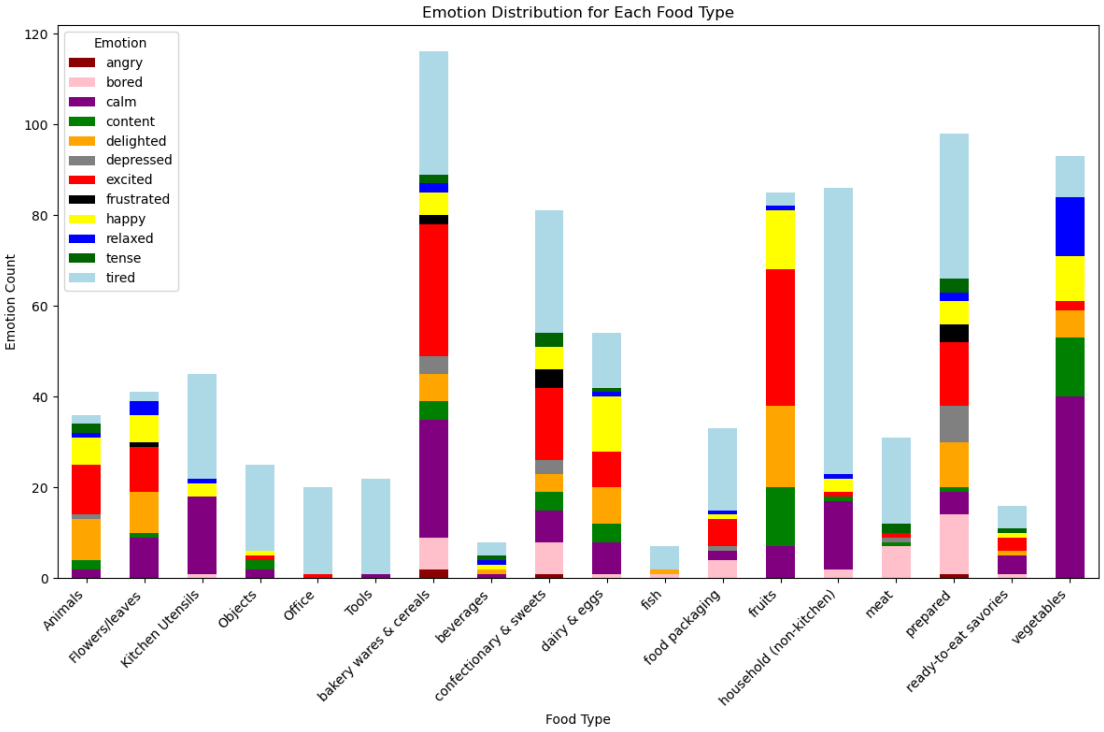


Figure 2.3: FoodPics Extended 2022 Dataset VA Distribution by Category

- bored: $0.33 < \text{slope} \leq 0.66$
- depressed: $0 < \text{slope} \leq 0.33$
- frustrated: $-0.33 < \text{slope} \leq 0$
- angry: $-0.66 < \text{slope} \leq -0.33$
- tense: $\text{slope} \leq -0.66$

This methodology groups samples with similar emotional characteristics together. It dedicated clear boundaries between different emotional states. The slope-based approach effectively captures the relationship between activation level (arousal) and emotional valence. Additionally, It was providing a systematic way to map continuous VA values to discrete emotion categories for analysis and visualization purposes.

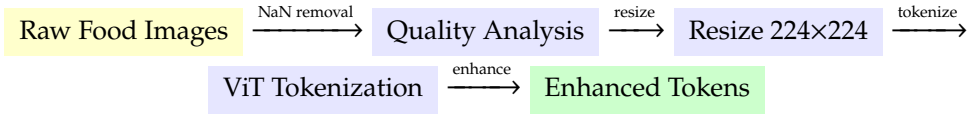
Food Category Analysis: The dataset contains WHO food categories and non-food categories, enabling emotion pattern analysis on different food types. Figure 2.3 illustrates the emotion distribution across food categories, revealing distinct emotional profiles for different food types. The analysis shows that bakery items and confectionery tend to excited emotions, and vegetables are

more associated with calm emotions. The nonfood items such as household items generally produce tired responses.

2.1.3 Data Processing Pipeline Overview

Below illustrates the complete data processing pipeline for both modalities, showing the systematic transformation from raw data to final token representations.

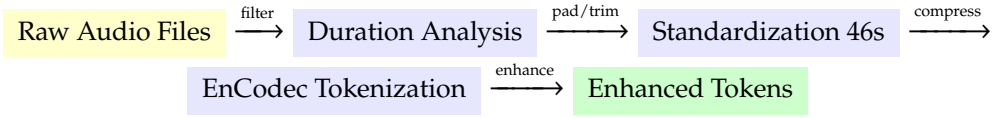
Image Processing Pipeline



Input: FoodPics Dataset (1,066 food images)
Process: Remove corrupted images, standardize to 224x224 pixels, extract patch embeddings
Output: Token sequences with or without valence-arousal values and modality flags

Figure 2.4: Image Processing Pipeline

Audio Processing Pipeline - Version 1-2
Direct EnCodec Approach



Input: DEAM Dataset (1,802 audio tracks)
Process: Filter tracks shorter than 46s, pad/trim to exact 46s duration, neural compression
Output: Compressed audio tokens with valence-arousal values and modality flags
Challenge: Different tokenization approach compared to image ViT tokens

Figure 2.5: Audio Processing Pipeline - Version 1-2

The Version 3 pipeline represents a significant advancement by enabling unified processing of both modalities through the same ViT architecture, which facilitates better cross-modal learning and shared representation development.

2.1.4 Data Representation and Tokenization

Both audio and image data has undergone tokenization processes to convert them into sequence representations suitable for the multimodal VAE architec-

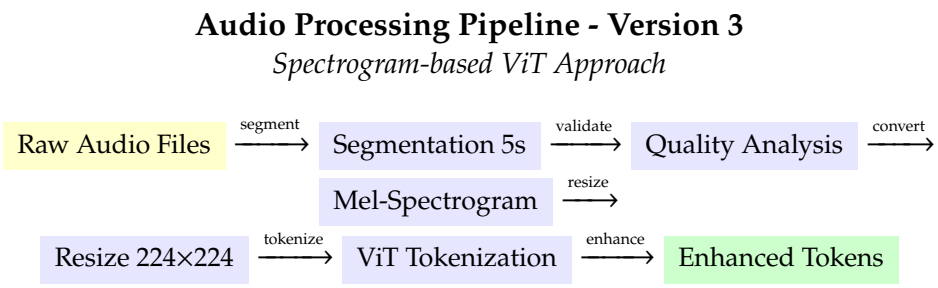


Figure 2.6: Audio Processing Pipeline - Version 3

ture. The tokenization approach evolved significantly between different versions to optimize model performance and enable effective cross-modal learning.

Tokenization Evolution Across Versions

Versions 1-2 (Mixed Tokenization):

- **Images:** Processed using a custom Vision Transformer configuration to produce 64 tokens, each with 2,048 dimensions
- **Audio:** Initially tokenized using the EnCodec model, producing 32 tokens with 3,450 dimensions each
- **Dimension Matching:** To enable unified processing through a single encoder, audio sequences were interpolated and projected to match the image token dimensions (64 x 2,048)

Version 3 (Unified ViT Approach):

- **Images:** Processed using standard ViT-base-patch16-224 configuration
- **Audio:** Mel-spectrograms processed using the same ViT architecture as images
- **Unified Processing:** Both modalities produce identical token structures (196 tokens x 768 dimensions)

Vision Transformer Configuration

For the final approach (Version 3), both modalities used the same ViT architecture:

- **Architecture:** ViT-base-patch16-224 (standard pre-trained configuration)
- **Input Size:** 224x224 pixels (for both images and mel-spectrograms)
- **Patch Size:** 16x16 pixels, resulting in 196 patches per input
- **Hidden Dimension:** 768
- **Output:** 196 patch tokens + 1 CLS token per sample

Token Enhancement with Contextual Information

Additional contextual information was integrated at the token level. This enables effective multimodal learning and emotion-conditioned generation:

Valence-Arousal Integration: The normalized VA values corresponding to each sample were replicated and appended to every token in the sequence. This ensures that emotion information is available at each sequence position during processing.

Modality Identification: Binary modality flags were added to distinguish between image tokens (flag = 0) and audio tokens (flag = 1). This allows the model to learn modality-specific processing patterns while maintaining shared latent representations.

Final Token Dimensions:

- **Version 3:** Both modalities: 771 dimensions (768 ViT features + 2 VA values + 1 modality flag) and 768 dimensions (768 ViT features) + 3 dimensions (2 VA values + 1 modality flag)
- **Versions 1-2:** Image tokens: 2,051 dimensions (2,048 features + 2 VA + 1 modality); Audio tokens: 3,453 dimensions (3,450 features + 2 VA + 1 modality)

2.1.5 Dataset Splits and Training Configuration

The final preprocessed datasets were divided using an 80-20 split ratio, applied consistently across both modalities:

Training Set: 80% of the data

- Approximately 853 images from FoodPics
- Approximately 10,379 audio chunks from DEAM segmentation for version 3 and 1395 audio for version 1-2

Testing/Validation Set: 20% of the data

- Approximately 213 images from FoodPics
- Approximately 2,595 audio chunks from DEAM segmentation for version 3 and 349 audio for version 1-2

The splitting strategy ensured balanced representation of emotion ranges in both training and testing sets, preventing bias toward specific emotional categories or ranges.

2.1.6 Data Quality Control and Validation

Quality control methods were comprehensive to ensure dataset integrity. These methods were applied throughout all processing stages:

ID-Based Tracking System

Each sample was assigned unique identifiers that were tracked through all processing stages. CSV files were maintained for both VA values and processed tokens, enabling verification of data consistency and completeness at each processing step.

Token Validation

Tokenization processes included logging mechanisms to identify and handle failed samples. The validation process included:

- **Dimensional Consistency:** All tokens were verified for correct dimensions within each version
- **Corruption Detection:** Failed tokenization attempts were logged and reprocessed to maintain dataset completeness
- **Missing Data Handling:** Systematic checks identified and addressed any missing associations between tokens and emotion annotations

Cross-Validation Checks

Regular validation was performed to ensure:

- **VA Range Validation:** Emotion values were checked for proper normalization and range compliance
- **Modality Flag Consistency:** Correct assignment of modality identifiers across all samples
- **Token-Annotation Alignment:** Verification that each token sequence correctly corresponds to its emotion annotations

Processing Pipeline Validation

Each major processing step was validated:

- **Audio Length (Version 1-2):** Verification that all 46 second audios were properly extracted and contained valid audio data
- **Audio Segmentation (Version 3):** Verification that all 5 second chunks were properly extracted and contained valid audio data
- **Mel-Spectrogram Generation:** Confirmation that all spectrograms were successfully generated and properly formatted
- **Image Resizing:** Validation that all images were correctly resized to 224x224 without corruption

Quality control measures resulted in a robust, validated dataset suitable for training the multimodal VAE system. The dataset maintains integrity of emotion annotations while ensuring technical compatibility across different model architectures and processing approaches.

2.1.7 Dataset Statistics Summary

Final Dataset Composition:

- **Images:** 1,066 food images with complete emotion annotations
- **Audio:**1744 forty-six second audios with static annotations for version 1-2
- **Audio Chunks:** 12,974 five-second audio segments with dynamic emotion annotations for version 3
- **Emotion Coverage:** Full valence-arousal space coverage with balanced representation

- **Token Dimensions:** 2,051 dimensional tokens for image modality, 3,453 dimensional tokens for audio modality in version 1-2. Unified 771 dimensional and 768 dimensional tokens for both modalities in version 3
- **Train/Test Split:** 80/20 ratio maintaining emotion distribution balance

This dataset configuration provides a solid foundation for training the multimodal VAE system. It provides to learn cross-modal emotion-based relationships and generate coherent audio content from visual emotion.

3

Material and Methodologies

3.1 Methodologies

3.1.1 Token-Based Representation

In modern deep learning architectures, it is especially important for multi-modality tasks. Token-based representation has emerged as a fundamental approach for processing different types of data within unified framework. A token represents a discrete unit of information that encapsulates meaningful features from the input data. This allows neural networks to process sequential information efficiently [21].

For multimodal systems, tokens serve as a common representational currency that allows different modalities to be processed within the same architectural framework [6]. When it comes to image-to-audio generation, audio tokens encode the temporal and spectral properties of sound, whereas visual tokens extract spatial and semantic information from images. This unified representation enables the model to learn cross-modal relationships and perform meaningful translations between modalities.

The token to token paradigm offers several advantages over traditional approaches [9]:

1. It enables sequence-to-sequence modeling across different modalities
2. It allows for variable-length outputs regardless of input size

3. It facilitates the application of attention mechanisms and transformer architectures to multimodal problems

This approach is particularly beneficial for cross-modal generation tasks where the output modality may have different temporal or spatial characteristics than the input.

Figure 3.1: Token sequence example showing visual patches converted to tokens and corresponding audio token sequences.

3.1.2 Variational Autoencoders for Multimodal Learning

Variational Autoencoders (VAEs) represent a significant advancement over traditional deterministic autoencoders by introducing probabilistic modeling into the latent representation [22]. VAEs learn a probability distribution over the latent space, allowing for controlled generation and improved regularization in contrast to traditional autoencoders that map inputs to fixed latent codes.

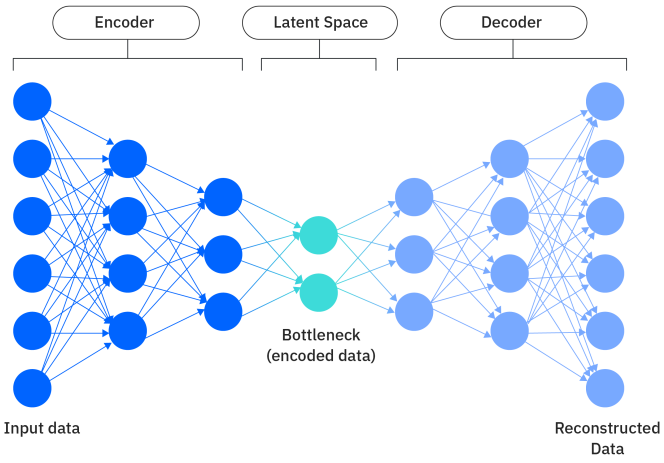


Figure 3.2: VAE architecture diagram showing encoder-decoder structure

Theoretical Foundation

The VAE framework consists of two primary components: an encoder network $q_\phi(z|x)$ that approximates the posterior distribution of latent variables z given input x, and a decoder network $p_\theta(x|z)$ that reconstructs the input from the latent representation. The training objective combines reconstruction accuracy with regularization through the Evidence Lower Bound (ELBO):

$$\mathcal{L}(\theta, \phi; x) = \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] - D_{KL}(q_\phi(z|x)||p(z)) \tag{3.1}$$

where D_{KL} represents the Kullback-Leibler divergence that encourages the learned posterior to remain close to a prior distribution, typically a standard Gaussian [23].

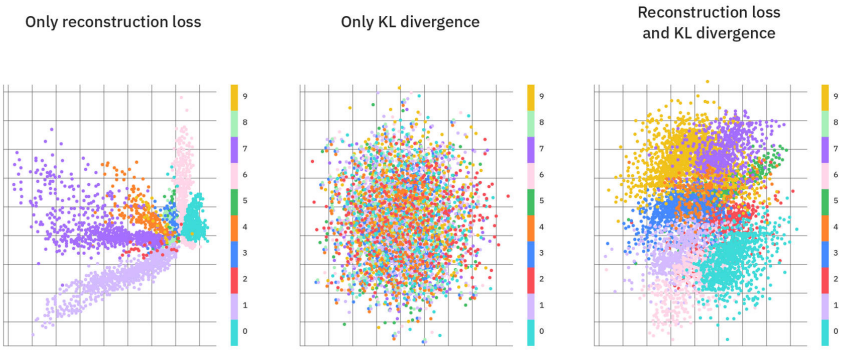


Figure 3.3: KL divergence regularization.

VAEs offer several key advantages for multimodal learning tasks compared to deterministic approaches:

1. **Continuous Latent Space**: Probabilistic latent space enables smooth interpolation between different modalities and it supports meaningful latent space arithmetic [24].

2. **Regularization**: The KL divergence term acts as a regularizer, preventing overfitting and encouraging. The model for learning generalizable representations across modalities.

3. **Uncertainty Modeling**: The probabilistic nature allows the model to capture and express uncertainty in cross-modal mappings. That is crucial when there are multiple valid outputs for a single input [25].

4. **Disentangled Representations**: VAEs naturally encourage disentangled latent representations. It enables better control over the generated outputs and improved interpretability [26].

The reparameterization trick ($z = \mu + \sigma \odot \epsilon$, where $\epsilon \sim \mathcal{N}(0, I)$) enables backpropagation through stochastic latent variables. That makes end-to-end training feasible while maintaining probabilistic benefits [22].

3.1.3 Residual Blocks

Residual blocks, introduced by [27], address the fundamental challenge. That is, for training very deep neural networks by mitigating the problem of vanishing gradients. The core innovation lies in the introduction of skip connections that allow information to flow directly across multiple layers.

Mathematical Formulation

A residual block implements the transformation:

$$y = \mathcal{F}(x, \{W_i\}) + x \tag{3.2}$$

where x represents the input, $\mathcal{F}(x, \{W_i\})$ denotes the residual mapping learned by the stacked layers, and the addition operation performs element-wise addition. This formulation transforms the learning objective from mapping $\mathcal{H}(x)$ to learning the residual $\mathcal{F}(x) = \mathcal{H}(x) - x$ [27].

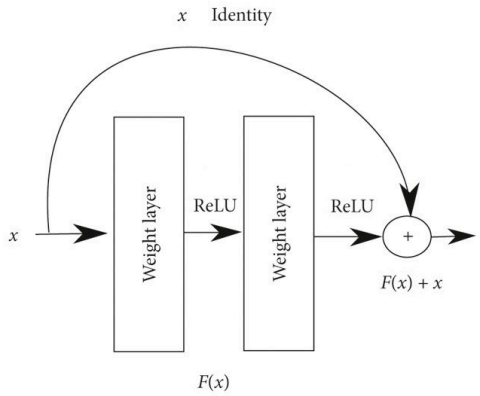


Figure 3.4: Residual Blocks diagram

Benefits for Deep Architectures

Residual blocks provide several crucial advantages:

1. **Gradient Flow**: Identity mapping provides a direct path for gradients during backpropagation. It enables training of networks with hundreds of layers without degradation. [27].

2. **Feature Reuse**: Lower-level features are directly accessible to higher-level layers, enabling efficient reuse of features and reducing the number of parameters needed [28].

3. **Representation Learning**: The residual formulation encourages the network to learn incremental refinements rather than complete transformations. It leads to more stable training dynamics [29].

Residual blocks are especially valuable in the context of multimodal VAEs because they make it possible to build deep encoder and decoder networks that can learn intricate cross-modal mappings while preserving training stability.

3.1.4 Skip Connections in Deep Architectures

Skip connections in our architecture extend beyond simple residual blocks by connecting non-adjacent layers within the same processing path. Unlike U-Net style architectures that connect encoder and decoder layers, our implementation uses intra-path skip connections to enhance gradient flow and feature reuse within both encoder and decoder networks separately.

Intra-Path Skip Connections

The skip connections in our architecture connect layers within the same processing path (either encoder or decoder). It allows information from previous layers to directly affect subsequent layers. Important features that could be lost due to numerous transformations are preserved by this design.

The mathematical formulation for our skip connections can be expressed as:

$$X_i = \mathcal{F}_i(\text{input}_i) + \mathcal{A}(X_{i-k}) \tag{3.3}$$

where X_i represents the output of layer i (either encoder or decoder), $\mathcal{F}_i$ is the layer transformation, $\mathcal{A}$ is an adaptation function for dimensional compatibility, and X_{i-k} represents the output from a previous layer k steps back.

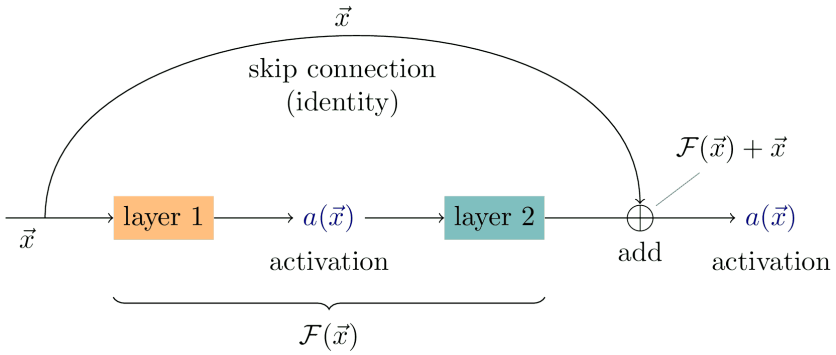


Figure 3.5: Skip connection diagram

3.1.5 Valence-Arousal (VA) Prediction Methodology

In this study, the "Valence-Arousal model" is integrated for emotion representation. The VA model defines emotions within a two-dimensional framework, valence (positive-negative emotional evaluation) and arousal (activation level). The integration of VA prediction into the methodology provides several critical advantages:

Emotion-Aware Cross-Modal Generation

VA values ensure that the model maintains emotional consistency during cross-modal generation. This method allows the emotional characteristics of the source modality to be reflected in the generated content. By conditioning the generation process on predicted emotional states, the model can create outputs that are not only structurally sound but also emotionally appropriate.

Quantitative Emotion Measurement

The VA space enables the transformation of subjective emotional experiences into objective, quantitative measurements. This facilitates the evaluation of model performance using emotion-aware metrics. The model is able to capture subtle emotional differences thanks to the continuous nature of the VA field, and thus can also capture subtle emotional differences that discrete categorical approaches might miss.

Latent Space Regularization

Latent representations are forced to learn emotion-related structures by the VA prediction task. More meaningful representations result from this auxiliary task, which also enhances the latent space's semantic organization. Emotional supervision makes the latent space more interpretable and structured around emotionally relevant features.

Two-Phase Training Strategy

The proposed methodology implements a two-phase training strategy for emotion learning:

1. **Phase 1 - Real VA Learning**: The model learns emotion representation using ground truth VA values. This phase ensures stable learning and accurate capture of emotional ground truth. During this phase, the model focuses on establishing reliable mappings between input features and emotional states.
2. **Phase 2 - Predicted VA Fine-tuning**: The model develops autonomous emotion prediction capability using its own predicted VA values. This transition prepares the model for real-world application scenarios where ground truth emotional labels may not be available.

This progressive training approach allows the model to first establish a solid foundation in emotion recognition before transitioning to autonomous prediction, leading to more robust and generalizable emotion-aware generation capabilities.

3.2 Model

In this section, variational autoencoder (VAE) based model architectures for the multimodal emotion based generation model are presented in detail. Two different approaches are adopted in this research: models using modality-specific tokenizers and models based on a single tokenizer. Both approaches aim to learn emotion representation in valence-arousal (VA) space and translate between modes.

The models aim to gain the ability to translate from one modality to another by representing image and audio data in a common latent space. For this purpose, VAE-based approaches based on encoder-decoder architecture have been developed and different training strategies have been tested.

Both approaches have its advantages and disadvantages. Throughout this project, which method was more sustainable was analyzed.

3.2.1 Modality-Specific Tokenizer Model

In this approach, specially trained tokenizers are used for each modality. Vision Transformer (ViT) based models for image data and transformer, EnCodec based models for audio data perform modality-specific feature extraction. The main advantage of this method is that it can optimally extract features from each modality.

The structure that generates the target modality from the latent space and VA values Within this approach, two different model versions have been developed: the basic VAE architecture (Version 1) and the advanced architecture (Version 2).

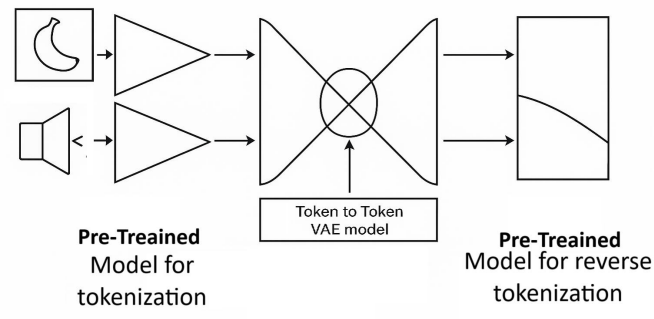


Figure 3.6: Model Diagram for Modality-Specific Tokenizer Model

Intermediate Fusion Strategy The architectural choice described above represents an intermediate fusion (or feature-level) strategy, which is the foundation for this approach. In these models, fusion occurs within the shared VAE architecture after modality-specific features have been extracted and dimensionally adapted. This design aims to preserve the unique characteristics of each modality by processing them with specialized tokenizers (ViT and EnCodec) before merging them in a common latent space.

Addressing Dimensional Differences in Modalities

Each modality was tokenized using its own specific tokenizer, which created significant dimensional mismatches at the initial stage. The project aimed to

implement a single encoder architecture in the model. The main reason for this approach was to enable he model to work with tokenized data from any tokenizer (regardless of whether the modality is image, audio, or other types), allowing adaptation to any desired data type. This goal aimed to enhance the model's generalizability and eliminate the need to develop separate models for different data types.

Initial Dimensional Differences Research was conducted on how to handle the different dimensions, as the size differences between modalities could not be ignored. The outputs obtained without adjusting the tokenizers were `[196, 768]` for images and `[32, 3450]` for audio data. This dimensional difference created approximately 4-5 times disproportion in both sequence length (196 vs 32) and feature dimensions (768 vs 3450).

Data Loss Minimization Strategy No additional neural network operations were applied to minimize data loss. The main reason for this approach was that using untrained additional layers could cause the network to produce inefficient outputs and negatively affect the model's learning process. Furthermore, it would deviate from the goal of keeping the model simple, which had been maintained from the beginning. This decision was made to control model complexity in the initial phase and minimize random factors that could negatively impact the learning process.

Vision Transformer Adaptation After research, it was decided to adjust the Vision Transformer (ViT) model used for image tokenization and modify its outputs. This approach aimed to obtain the desired dimensional output by changing the structural parameters of a pre-trained model.

The modification began with changing the patch size and hidden dimension parameters:

- **Patch Dimension Adaption:**
  - Original patch size: 16×16 pixels
  - New patch size: 28×28 pixels
  - With this change, the number of patches obtained from a 224×224 image decreased from (14×14) to 64 (8×8)

- **Hidden Dimension Expansion:**

  - Original hidden dimension: 768
  - New hidden dimension: 2048
  - This increase aimed to preserve the richness of feature representation and obtain a dimension more compatible with the audio modality

After these operations, the image token size changed to `[64, 2048]`. This approach provided these advantages:

- **Preserving pre-trained knowledge**: The visual representations learned by ViT have been preserved.
- **Minimal intervention**: Only structural parameters were changed, weights were preserved
- **Approaching audio modality**: The sequence length (64) became closer to the audio modality

EnCodec Limitations and Audio Modality Adaptation Unfortunately, it was not possible to make similar changes to the EnCodec model or implement suitable modifications to achieve the desired result. Since EnCodec's architectural structure and training process did not support the desired dimensional changes, a different adaptation strategy was required for the audio modality.

Feature Dimension Harmonization: The primary problem was the dimensional mismatch between audio features (3450) and image features (2048). A learnable linear projection layer was used to map audio features to the target image feature space. This approach mapped 3450-dimensional audio features to 2048 dimensions.

This approach offered several advantages over alternative methods like PCA or feature selection:

- **Learnable transformation**: Projection parameters are optimized during training to preserve semantically relevant information
- **Gradient flow**: Enables end-to-end optimization of dimensional mapping

Sequence Length Normalization: The sequence length difference (32 vs 64 tokens) was addressed through bilinear interpolation. This method provided smooth temporal resampling while maintaining sequential dependencies.

Bilinear interpolation was chosen over simpler alternatives (padding, truncation, or repetition) for the following capabilities:

- Ensuring temporal consistency in feature representations
- Providing smooth transitions between adjacent time steps
- Preserving the overall temporal structure of the audio sequence

Unified Token Representation After the adaptation process, all modalities became compatible with the standardized token format: `[batch_size, 64, 2051]`. Each token encoded the following:

- **2048-dimensional features**: Harmonized cross-modal representations
- **2-dimensional VA values**: Valence and arousal annotations
- **1-dimensional modality indicator**:Binary flag distinguishing between modalities

This unified representation enabled the downstream VAE architecture to process heterogeneous inputs through shared encoder and decoder networks. This facilitates effective cross-modal learning and generation capabilities.

Hybrid Adaptation Approach As a result, a hybrid adaptation approach was adopted:

1. **Image modality**: Adaptation to `[64, 2048]` format by changing ViT structural parameters
2. **Audio modality**: Mapping to `[64, 2048]` format using learnable linear projection and bilinear interpolation

This approach succeeded in both minimizing data loss and keeping model complexity under control by using the most suitable adaptation strategy for each modality.

Table 3.1: Cross-modal dimensional adaptation process

Modality	Original Size	Adaptation Method	Final Size
Image	[196, 768]	ViT parametric change	[64, 2048]
Audio	[32, 3450]	Linear projection + bilinear interpolation	[64, 2048]

3.2.2 Single Tokenizer-Based Model

This approach aims to achieve a more homogeneous feature representation by using a common tokenizer which is Vision Transformer (ViT) based models for all modalities. This method aims to create a common representation space by better capturing similarities between modalities.

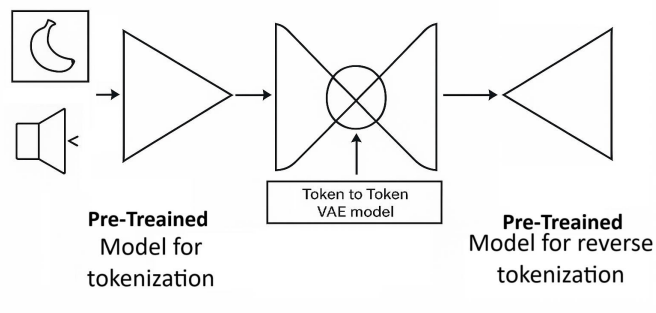


Figure 3.7: Model Diagram for Single Tokenizer-Based Model

This approach aims to learn a more consistent representation across modalities and improve translation quality. This approach eliminates the token dimension restrictions and facilitates a more flexible encoder design.

Early Fusion Strategy: This architecture is an example of an early fusion (or data-level) strategy and is implemented in modals. By converting audio into a visual spectrogram format during preprocessing, fusion is achieved at the data level before the model is even engaged. This allows a single, unified ViT architecture to process both modalities, drastically simplifying the cross-modal learning task.

Riffusion for Audio-to-Visual Representation

This research employs Riffusion, a method that enables the application of Vision Transformer (ViT) architectures to audio data processing [30]. The technique transforms audio signals into visual spectrograms, making them compatible with image-based neural network architectures. In this approach, audio waveforms are converted into mel-scale spectrograms, which is allowing the same ViT framework used for visual data to process audio inputs effectively.

The conversion process preserving the temporal and frequency characteristics of audio signals while representing them in a visual format that can

be tokenized within the same [196, 768] dimensional space as image tokens. By implementing Riffusion's audio-to-spectrogram transformation, this study achieves a unified tokenization strategy where both image and audio modalities share identical dimensional representations. This approach eliminates the requirement for modality-specific adaptation techniques, such as previously mentioned in modality-specific tokenizer model part. In addition, it facilitates more effective cross-modal learning within the VAE framework. The resulting architecture demonstrates improved consistency across different data types. This approach enhances the model's ability to learn shared representations between visual and auditory information.

Table 3.2: Modality Dimension Representation

Modality	**Adaptation Method**	**Size**
Image	Vision Transformer (ViT)	[196, 768]
Audio	Vision Transformer (ViT)	[196, 768]

3.3 Modality-Specific Tokenizer Model Trials

3.3.1 Version 1

Encoder Architecture and Functionality

The encoder component handles the transformation of input data into latent space and has a three-layered hierarchical structure. The first layer reduces the 2051-dimensional input features to 1024 dimensions. The second layer processes these features along with the original input data to produce 512-dimensional representations. The final layer generates the final representation in 256 dimensions. This hierarchical approach enables the model to learn features at different levels of abstraction.

Each encoder layer consists of advanced components that reflect modern deep learning practices. Linear transformations are followed by Layer Normalization. GELU activation functions, and Dropout regularization uses to prevent overfitting. Additionally, specially designed Residual Blocks are included at the end of each layer to improve gradient flow.

One of the most important features of the encoder is the skip connection mechanism. These connections provide direct information flow between differ-

ent layers, reducing the gradient vanishing problem and enabling more effective feature learning. Skip connection adapters apply linear transformations to resolve dimension mismatches, ensuring lossless information transfer.

Latent Space Design and Reparameterization

The 256-dimensional features obtained from the encoder's final layer are processed in two separate branches for latent space projection. These branches produce mean (mu) and log variance ($\log \sigma^2$) values respectively. This design enables the probabilistic latent representation that is fundamental to VAEs.

Using the reparameterization trick, the latent vector is obtained as a stochastic representation sampled from a normal distribution rather than a deterministic mean value. This approach increases the model's generation capacity while providing regularization effects. During training, sampling is performed, while during inference, deterministic results can be obtained using only mean values.

The reparameterization trick is a critical mathematical innovation that enables VAE's to be trained using backpropagation. Traditionally, gradient flow through stochastic nodes is not possible because sampling operations are not differentiable. This problem arises when derivatives of the latent variable z with respect to μ and σ The essence of the trick involves converting stochastic sampling into deterministic transformation. This reparameterization allows gradients to smoothly flow to the parameters μ and σ while preserving stochastic properties.

VA Predictor Module

The VA Predictor, which estimates valence-arousal values from latent space, is designed as a four-layer deep neural network. This module handles the task of extracting emotional information from latent representations, processing through progressively decreasing dimensions (256→128→64→2). The use of Tanh activation in the final layer guarantees outputs within the [-1,1] range.

In the design of the VA Predictor, Layer Normalization and Dropout are applied between each layer to ensure model stability. The success of this module directly affects the quality of cross-modal generation because the predicted VA values are used as conditioning information in the decoder.

Decoder Architecture and Conditional Generation

The decoder has a symmetrical structure to the encoder but is specially designed for conditional generation. It takes a concatenated combining latent vector, VA values, and modality indicator, using this information to perform reconstruction in the target modality. This approach enables the model to perform controlled generation in different modalities and different emotional states.

The three-layer structure of the decoder works with expanding dimensions (latent_dim+3→512→1024→2051), reverse to the encoder. Skip connections are used in each layer to optimize the gradient flow similar to the encoder. This symmetrical design minimizes information loss resulting from the information bottleneck.

Two-Phase Training Strategy

The model achieves optimal performance using a two-phase training strategy. In the first phase (Real VA Training), ground truth VA values are used in the decoder, allowing the model to learn reconstruction with real emotional labels. During this phase, the VA predictor is also trained in parallel to learn extracting emotional information from latent space.

In the second phase (Predicted VA Fine-tuning), the predicted VA values are started to be used in the decoder. This transition enables the model to perform completely autonomous generation. The learning rate is reduced in this phase to achieve fine-tuning characteristics. This two-phase approach optimizes both reconstruction and generation capabilities of the model.

The fundamental motivation of the two-phase training approach is to develop the model's autonomous VA-guided generation capability. The final objective is for the model to analyze the emotional content of any given input, predict corresponding VA values, and perform guided generation with this emotional information.

First Phase: Training with Real VA Values: Ground truth VA values are used as conditioning in the decoder during this phase. This supervised approach enables the model to learn optimal reconstruction quality while the VA predictor module learns to extract emotional information from latent space in parallel. This phase answers the question "how to perform generation under perfect conditioning."

Second Phase: Fine-tuning with Predicted VA Values: In the second phase,

the decoder begins using predicted VA values. This transition makes the model completely self-sufficient. It can now perform generation with automatically extracted emotional information from input without requiring external VA annotation. Learning rate reduction provides fine-tuning character, preventing disruption of already learned representations.

This two-phase methodology provides smooth transition from supervised learning to self-supervised learning, optimizing both reconstruction and autonomous generation capabilities of the model.

Loss Function and Optimization

The model optimizes a weighted combination of four different loss components. The total loss function is formulated as follows:

Total Loss:

$$\mathcal{L}_{total} = \mathcal{L}_{recon} + \beta \times \mathcal{L}_{KL} + \lambda_{VA} \times \mathcal{L}_{VA} + \mathcal{L}_{L2} \tag{3.4}$$

Mean Squared Error (Reconstruction Loss): This loss minimizes the difference between original and reconstructed data. MSE provides detailed reconstruction at token level, enabling the model to learn input-output mapping:

$$\mathcal{L}_{recon} = \frac{1}{N} \sum (x_{reconstructed} - x_{original})^2 \tag{3.5}$$

KL Divergence Loss: The regularization component of VAE, KL divergence, brings the learned posterior distribution of latent space closer to standard normal distribution. This ensures the latent space is smooth and interpolatable:

$$\mathcal{L}_{KL} = -\frac{1}{2N} \sum (1 + \log(\sigma^2) - \mu^2 - \sigma^2) \tag{3.6}$$

VA Prediction Loss: This loss minimizes the difference between valence-arousal values predicted from latent space and ground truth values, ensuring emotional information is preserved in latent representation:

$$\mathcal{L}_{VA} = \frac{1}{N} \sum (VA_{predicted} - VA_{true})^2 \tag{3.7}$$

L2 Regularization Loss: This weight decay mechanism prevents model pa-

rameters from taking excessively large values, reducing overfitting:

$$\mathcal{L}_{L2} = \lambda_{L2} \sum ||\theta||^2 \tag{3.8}$$

Dynamic adjustment of loss weights enables focus on different objectives at different stages of the training process. The loss of VA weight (λ_{VA}) is aimed at developing the emotion prediction capacity of the model.

AdamW optimizer is used for optimization, and the training is stabilized with Cosine Annealing with the Warm Restarts learning rate scheduler. This scheduler provides an an opportunity to escape local minima through periodic restarts.

Model Persistence and Continuous Training

A comprehensive checkpoint system has been developed for the continuity and reproducibility of model training. The model state, optimizer state, scheduler state, training history, and hyperparameters are saved at the end of each epoch. This system enables continuation from where the training left off if interrupted at any point.

Checkpoint files are automatically saved both at regular intervals and upon validation loss improvement (best model). Training resume functionality enables the model to continue training from exactly the same state by loading saved states. This feature is critically important for long-duration training processes and allows experiments with different hyperparameters.

Regularization and Stabilization Techniques

A comprehensive regularization and stabilization framework has been developed for the model to achieve robust and generalizable learning. The combination of these techniques prevents both overfitting and increases training stability in complex multi-modal learning tasks.

Structural Regularization Techniques

Dropout Regularization: Dropout deactivates randomly selected neurons during training, and preventing the model from becoming overly dependent on specific neurons. This technique solves the co-adaptation problem, enabling

each neuron to learn more independent and robust features. It is particularly important in multi-modal settings because it guarantees balanced learning of information from different modalities.

Layer Normalization: Layer Normalization is applied after each linear transformation. This technique stabilizes the training process by solving the internal covariate shift problem. Particularly in VAE architecture, it optimizes gradient flow in the deep structure of encoder and decoder and increases convergence speed. Provides more consistent performance compared to batch normalization due to the independence of batch size.

Residual Connections: Residual blocks are used in each encoder and decoder layer. These connections solve the gradient vanishing problem, enabling effective training in deep networks. By adding identity mapping between two sequential linear transformations, they allow the model to preserve simple mappings while learning complex transformations. This is particularly critical in multi-modal VAE as it minimizes information loss from information bottleneck.

Training-Level Stabilization Mechanisms

Gradient Clipping: This technique prevents gradient explosion problems, increasing the stability of training. It is particularly important in VAE training because the optimization landscape of KL divergence loss can sometimes be unstable. Gradient clipping provides smooth convergence by softening large gradient spikes.

Weight Decay (L2 Regularization): This regularization increases generalization capability by preventing model parameters from taking excessively large values. In cross-modal generation tasks, it ensures that mapping between different modalities is smooth and meaningful.

Early Stopping: Training is automatically terminated when no improvement is observed in validation loss for previously defined epochs. This mechanism detects the point where overfitting begins, capturing optimal generalization performance.It is especially valuable in multi-modal learning because complex models can quickly overfit.

Advanced Optimization Strategies

Learning Rate Scheduling: Cosine Annealing with Warm Restarts scheduler is used. This scheduler periodically decreases and increases learning rate according to cosine function, providing opportunity to escape local minima. Previously defined epochs for first restart and two times defined epochs for period doubling strategy are applied. This approach increases the chance of finding better global optimum in complex loss landscapes.

Warm up Strategy: Learning rate is increased linearly during the firsts epochs of training. This technique enables smooth transition from initial random weights to stable training regime. It carries critical importance particularly in large models and complex loss functions.

Multi-Modal Specific Stabilization

Balanced Sampling: Equal numbers of image and audio samples in each batch are guaranteed. This approach prevents modality bias, providing balanced learning. It is systematically maintained through "Balanced Batch Sampler" implementation.

Dynamic Loss Weighting: The VA loss weight is dynamically adjusted according to training progression. The gradual increase in phase 1 enables the gradual development of the VA prediction capability ($\lambda_{VA} = \min(1.0, initial\_weight + epoch \times 0.01)$). This adaptive weighting enables the balanced optimization of different loss components.

Latent Space Regularization and Disentangled Representation Learning

Latent space regularization is achieved with KL divergence loss and this regularization has a fundamental effect on VA prediction capability. KL loss brings learned posterior $q(z|x)$ closer to prior $p(z) = \mathcal{N}(0, I)$, ensuring latent space is structured and smooth.

The primary effect of this regularization on VA prediction is compact and meaningful organization of latent representation. The limitation of approaching standard normal distribution encourages samples with similar emotional content to cluster at nearby locations in latent space. This spatial organization enables the VA predictor to make consistent and robust predictions.

In terms of disentangled representation learning, the model implicitly attempts to separate content and emotion information in the latent space. The

dedicated architecture of the VA predictor promotes the specialization of specific dimensions of the latent vector to emotional attributes. This disentanglement enables independent control of content preservation and emotion transfer in cross-modal generation.

KL regularization ensures balanced optimization of the VA prediction loss, guaranteeing that the latent space is both expressive and interpretable. This dual objective enables the model to perform modality-cross transformations while preserving semantic meaning.

Cross-Modal Generation Mechanism

The model's cross-modal generation capacity is based on the principle of learning modality-independent representation in shared latent space. The generation process occurs in three stages: encoding from source modality to latent space, predicting emotional features from latent space, and decoding by combining target modality with conditioning. This approach enables cross-modal transformation while preserving semantic and emotional information.

Results

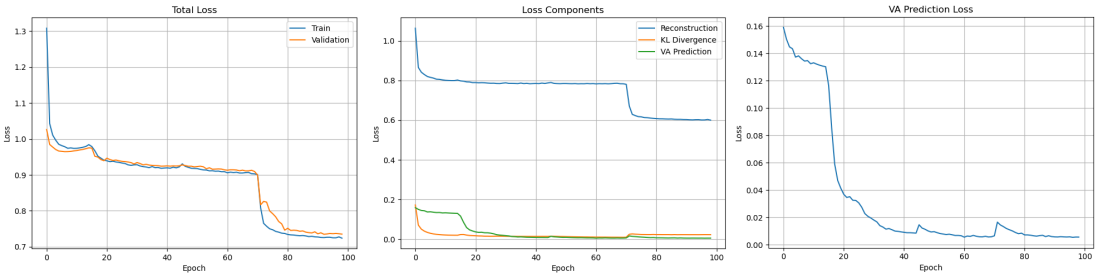


Figure 3.8: Graphic of Training Dynamics of Version 1

Graphic of Training Dynamics: Training shows a steady convergence with a steady decrease in loss. The fact that the difference between validation and training loss is not significant indicates that the model is not overfitting, but the high loss clearly shows that the model still needs improvement. The improvement in VA prediction loss indicates effective emotional learning.

Graphic of Generation Performance: The generated signals show significantly reduced amplitudes compared to the inputs, indicating limited reconstruction quality. While VA estimates remain within reasonable ranges, con-

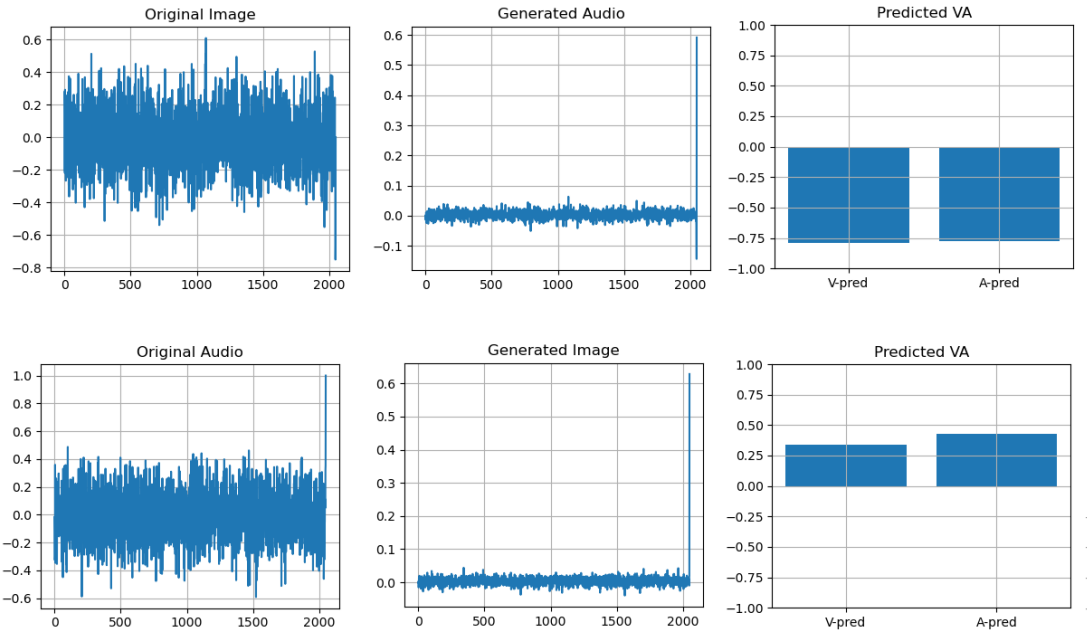


Figure 3.9: Graphic of Generation Performance of Version 1

servative output behavior indicates insufficient production accuracy, requiring architectural improvements in subsequent model iterations.

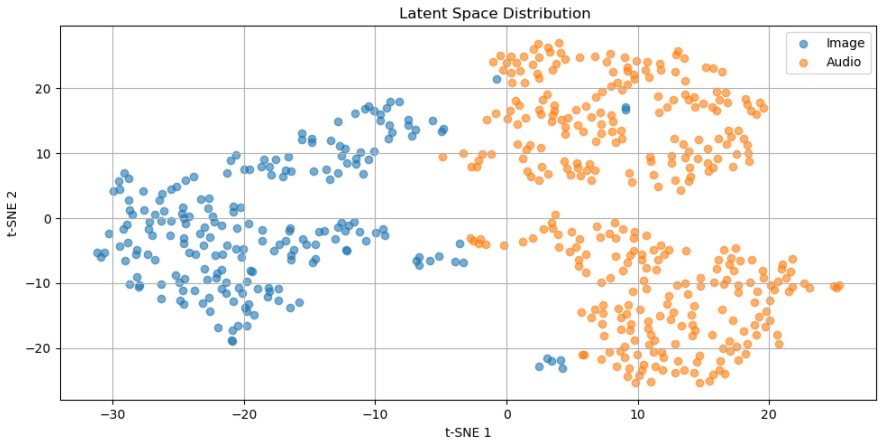


Figure 3.10: Graphic of Latent Space of Version 1

Graphic of Latent Space: By showing a clear distinction between distinct clusters and clear modality, it validates the model's ability to learn modality-specific representations. However, this strict distinction limits cross-modal feature sharing and semantic alignment between modalities. The latent space organization, while functionally sufficient, lacks the structure necessary for high-quality cross-modal synthesis.

3.3.2 Version 2

The improvements implemented during the transition from Model V1 to Model V2 have achieved significant advances in both architectural complexity and performance aspects of the model.

Architectural Improvements

The most crucial architectural modification involves the establishment of a deeper network structure by increasing the number of encoder and decoder layers from three to six respectively. This architectural enhancement follows a systematic dimensional reduction pattern: Layer_1(2048), Layer_2(1536), Layer_3(1024), Layer_4(768), Layer_5(512), and Layer_6(256). This deep architecture enables the model to learn more complex feature representations, while a skip connection system has also been developed to enhanced gradient flow.

The systematic redesign conducted in the skip connection architecture targets gradient flow optimization and long-range dependency preservation. These adapters maintain dimensional consistency while integrating LayerNorm, GELU activation, and Dropout regularization layers to optimize residual information flow. The skip connection topology mitigates the gradient vanishing problem through systematically defined dimensional mappings, and ensures information preservation that is critical in deep networks. This advanced skip architecture plays an essential role in preserving long-range temporal dependencies, especially in multi-modal contexts.

The Conditional Layer Normalization (CLN) system is one of the most innovative parts of Model V2. While Model V1 employed standard Layer Normalization exclusively, V2 generates distinct normalization parameters (γ and β) for each modality. This system operates through 64-dimensional modality embeddings and applies modality-specific normalization within each decoder layer. The CLN mechanism can be mathematically formulated as:

$$\text{CLN}(x, c) = \gamma(c) \cdot \text{LayerNorm}(x) + \beta(c) \tag{3.9}$$

where x represents the input features, c denotes the conditioning vector (modality embedding), and the modality-dependent parameters are computed as:

$$\gamma(c) = W_\gamma \cdot c + b_\gamma \tag{3.10}$$

$$\beta(c) = W_\beta \cdot c + b_\beta \tag{3.11}$$

Here, W_γ and W_β are learnable projection matrices, while b_γ and b_β represent bias terms. The LayerNorm operation is defined as:

$$\text{LayerNorm}(x) = \frac{x - \mu}{\sqrt{\sigma^2 + \epsilon}} \tag{3.12}$$

where μ and σ^2 are the mean and variance computed across feature dimensions, and ϵ is a small constant for numerical stability.

CLN significantly enhances cross-modal learning capacity by enabling the model to better comprehend the distinct statistical characteristics of visual and auditory modalities. This capability particularly improves model performance in cross-modal transition tasks and allows each modality to be processed according to its inherent dynamics.

In the training infrastructure, mixed-precision training support optimizes float16/float32 computations in CUDA environments and ensures numerical stability with a gradient scaling mechanism.

These comprehensive architectural modifications substantially strengthen Model V2's representational capacity and cross-modal learning ability, enabling superior performance achievement in multi-modal generation tasks.

Results

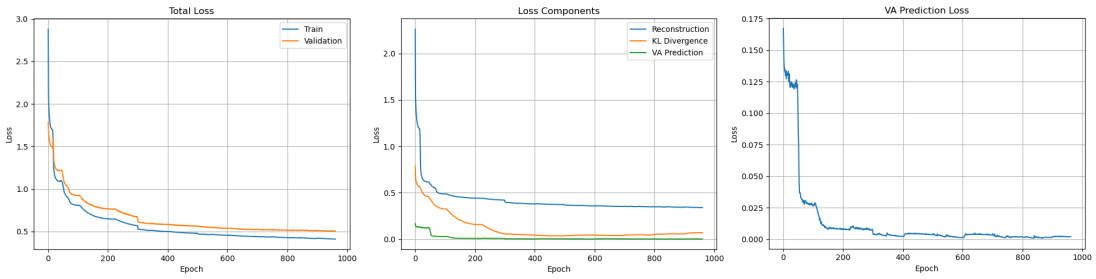


Figure 3.11: Graphic of Training Dynamics of Version 2

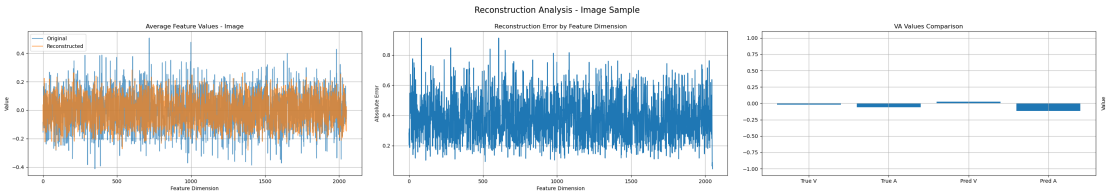


Figure 3.12: Graphic of Image Generation Performance of Version 2

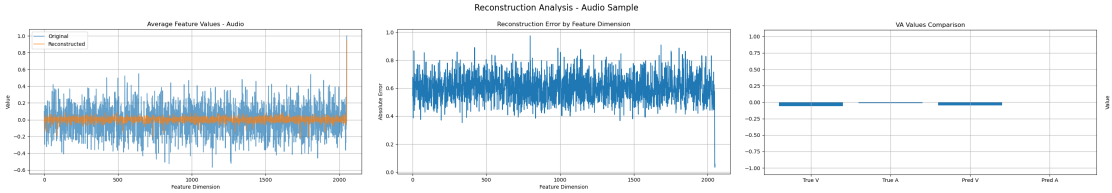


Figure 3.13: Graphic of Audio Generation Performance of Version 2

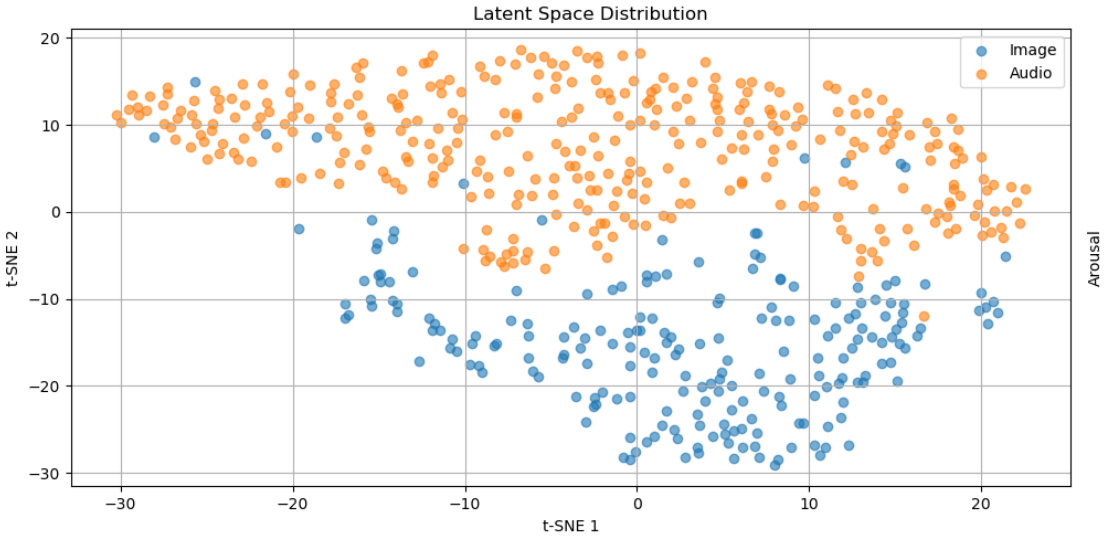


Figure 3.14: Graphic of Latent Space of Version 2

Examining of the outputs of Model V2, the decrease observed in loss values indicates that the model gradually improved during training in Figure 3.11. However, it is clear that the loss values are still well above the desired levels, which explains the poor quality of the production graph outputs, such as the generated examples in Figure 3.13 and Figure 3.12 show insufficient accuracy and consistency. It was especially in production tasks where the model struggles to maintain semantic consistency across modalities.

This performance analysis reveals that while Model V2 represents a significant advancement over its predecessor, substantial room for improvement persists. The difference between the current performance metrics and the targets indicates that architectural improvements and optimization strategies are necessary. Relatively high loss values are directly related to the quality limitations observed in the visual outputs and indicate that the model has not yet reached the desired level of representational coherence.

The valence-arousal prediction accuracy shows promising results with reasonable correlation coefficients in Figure 3.11, Figure 3.12 and Figure 3.13.

In conclusion, these results show that the Model V2 architecture needs to development and enhancement of the Model V2 architecture. The current results provide a basis for identifying specific areas for improvement and guide the direction of future model iterations.

3.3.3 Version 2.1

In this version, comprehensive improvements were implemented based on the performance analyses of the previous model (Version 2). Model Version 2.1 preserves the fundamental strengths of the base architecture, and introducing significant improvements in learning dynamics.

Architectural Improvements and Capacity Enhancement

One of the most significant architectural improvements in Model Version 2.1 is that involves expanding the latent space dimensionality from 256 to 384. This 50% capacity increase enables the model to learn richer and more detailed representations. In particularly yielding substantial improvements in cross-modal generation quality. This dimensional expansion facilitates more effective encoding of complex inter-modal relationships and enables more precise VA

(valence-arousal) predictions. The increase results growth in total parameter count and proportionally increases computational complexity.

Parallel to this capacity enhancement, the modality embedding dimension was elevated from 64 to 96. This 50% increase allows for enhanced representation of modality-specific conditioning information, and significantly improving the effectiveness of conditional normalization layers. This dimensional expansion enables more detailed capture of characteristic features across different modalities. So, it allows the model to make more precise distinctions between modalities.

Model Version 2.1 also introduces a fundamental enhancement in the conditional normalization mechanism through the Gated Conditional Layer Normalization (GCLN) approach. This normalization technique includes a gating mechanism that dynamically controls the impact of modality-specific transformations. Mathematically, GCLN is defined as:

$$\begin{aligned} \text{GCLN}(x, c) = \sigma(W_{gate} \cdot c + b_{gate}) \odot [\gamma(c) \cdot \text{LayerNorm}(x) + \beta(c)] \\ + [1 - \sigma(W_{gate} \cdot c + b_{gate})] \odot \text{LayerNorm}(x) \end{aligned} \tag{3.13}$$

where the modality-dependent parameters are computed as:

$$\gamma(c) = 1 + W_\gamma \cdot c + b_\gamma \quad \text{(scaling parameter)} \tag{3.14}$$

$$\beta(c) = W_\beta \cdot c + b_\beta \quad \text{(shifting parameter)} \tag{3.15}$$

$$\text{gate}(c) = \sigma(W_{gate} \cdot c + b_{gate}) \quad \text{(gating parameter)} \tag{3.16}$$

Here, W_γ, W_β, and W_{gate} represent learnable projection matrices, b_γ, b_β, and b_{gate} denote bias terms, σ is the sigmoid activation function, and c represents the modality embedding vector. The "+1" term in the $\gamma(c)$ computation ensures identity transformation at initialization. The gate mechanism is initialized with a bias value of -1, ensuring controlled application of conditioning effects during initial training phases.

Training Strategy Development

The loss function strategy in Model Version 2.1 underwent fundamental redesign. A Progressive MSE Loss methodology was implemented, enhancing learning dynamics by using different loss types at different stages of the training

process. Based on training progress ($p = \frac{epoch}{max\_epochs}$), the loss function is defined as:

$$L(x, \hat{x}) = \begin{cases} ||x - \hat{x}||_4^4 & \text{if } p < 0.3 \text{ (aggressive focus on large errors)} \\ ||x - \hat{x}||_2^2 & \text{if } 0.3 \leq p < 0.7 \text{ (standard MSE)} \\ ||x - \hat{x}||_1 & \text{if } p \geq 0.7 \text{ (fine detail optimization)} \end{cases} \tag{3.17}$$

In this progressive approach, the fourth-power loss (L4) employed in early phases exponentially penalizes large errors while relatively ignoring smaller errors. This encourages the model to initially focus on learning general structural characteristics. Standard MSE is utilized in intermediate phases, while L1 loss in final phases targets fine detail optimization. This progressive methodology, combined with increased latent capacity, hierarchically structures the model's learning process.

Learning Rate Optimization

A hybrid learning rate scheduler was developed to enhance training stability and expand model capacity. This system uses a two-step process that starts with cosine annealing after a predetermined epoch warm up period:

$$lr(t) = \begin{cases} lr_{base} \times \frac{t}{epoch} & \text{if } t \leq epoch \\ lr_{base} \times \cos\left(\frac{\pi(t-epoch)}{T-epoch}\right) & \text{if } t > epoch \end{cases} \tag{3.18}$$

The epoch warmup duration was determined to minimize the risk of gradient explosion caused by the expanded latent space during initial training phases. This formulation ensures stable training progression despite increased model capacity while enabling fine-tuning in later stages.

These enhancements, particularly the increased latent space capacity and advanced normalization mechanisms, enable Model Version 2.1 to demonstrate more stable training dynamics, improved modality adaptation, and superior cross-modal generation performance compared to its predecessor. Experimental results indicate that these developments provide significant improvements in the

model's overall performance and generalization capabilities.

Results

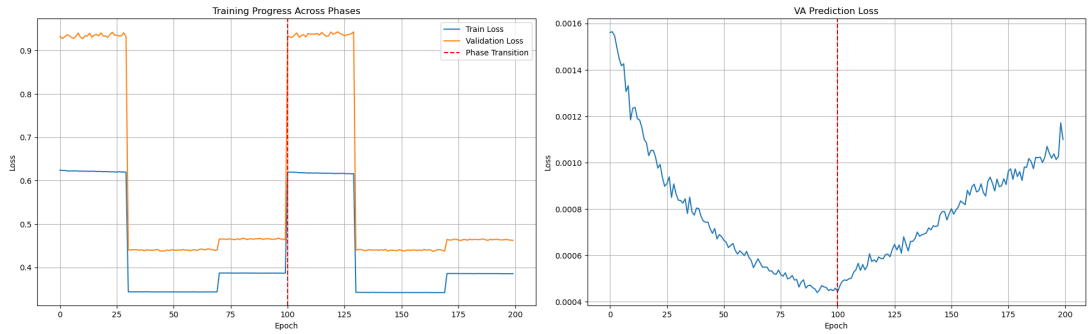


Figure 3.15: Graphic of Training Dynamics of Version 2.1

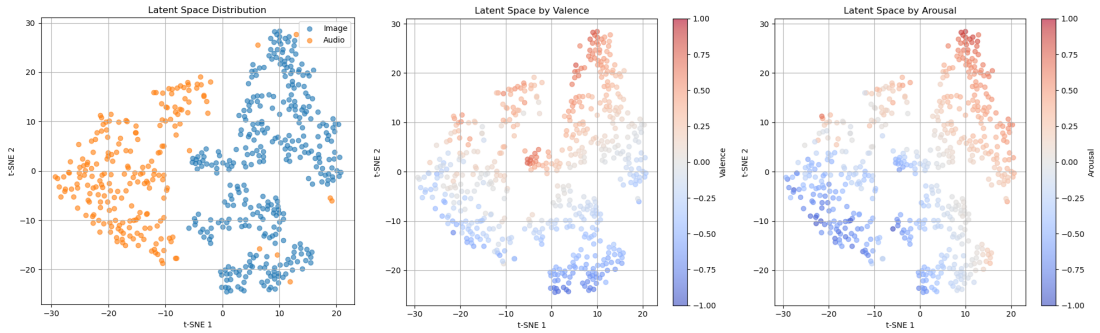


Figure 3.16: Graphic of Latent Space of Version 2.1

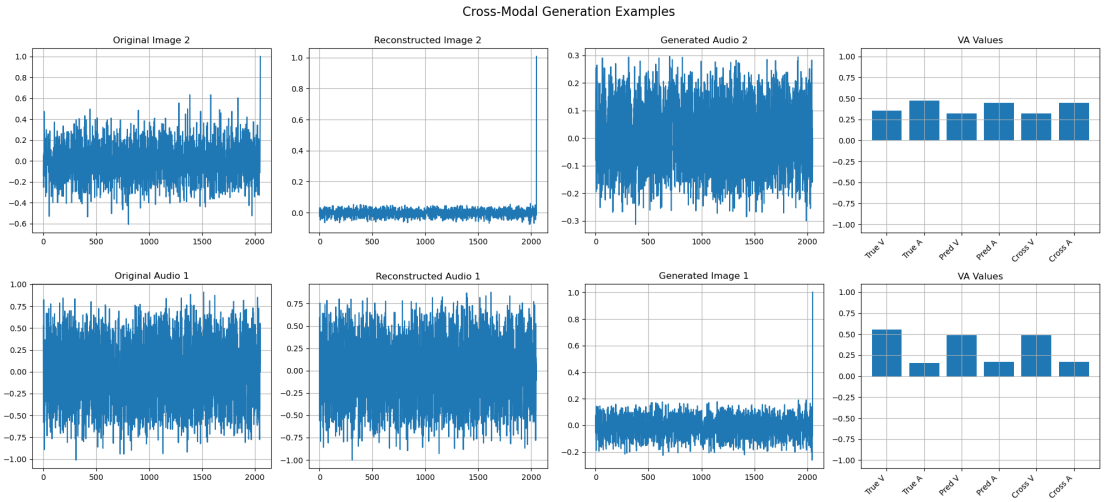


Figure 3.17: Graphic of Generation Performance of Version 2.1

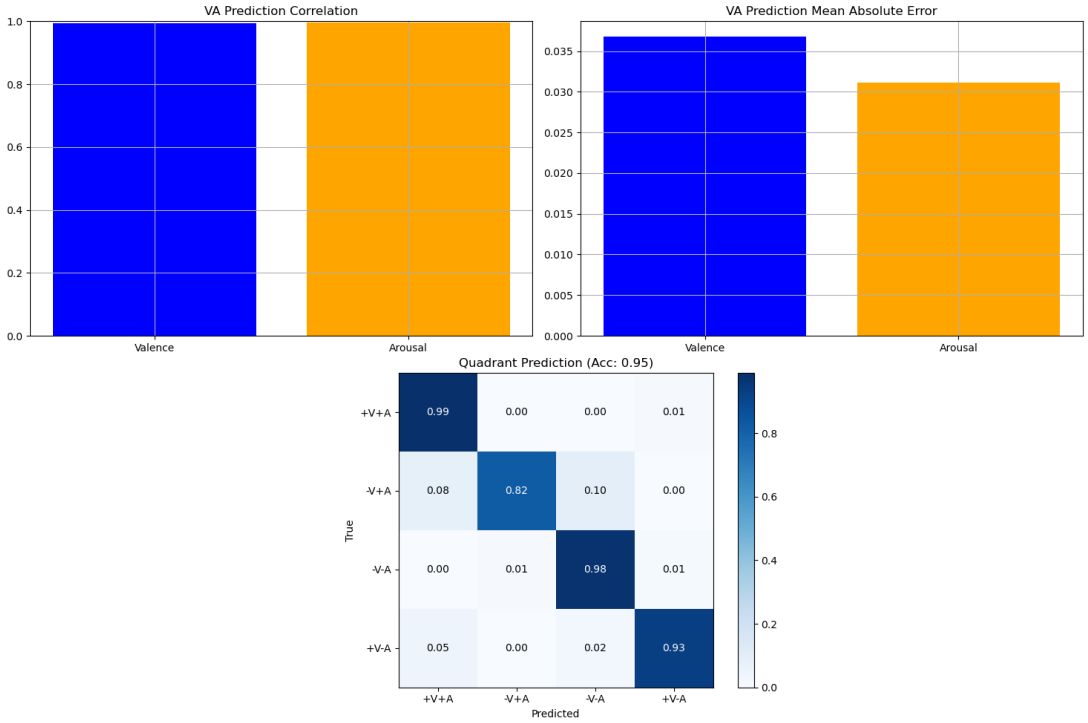


Figure 3.18: Graphic of Valence Arousal Performance of Version 2.1

When examining the model 2.1 results through the figures, the enhanced utility of the new GCLN becomes evident from the improved modality separation observed in the latent space graphical representations in Figure 3.16. Furthermore, the VA distribution demonstrates that the model has established a latent space organized according to both modality characteristics and VA values. However, analysis of the reconstruction loss in Figure 3.15 reveals that this performance remains insufficient, indicating that continued model development is necessary.

The generation in Figure 3.17 provides additional evidence of the insufficient reconstruction loss performance. The model's inability to comprehend semantic relationships, particularly in the visual modality, observed through the disparities in the generated outputs in Figure 3.16.

The VA prediction mechanism appears to deliver the desired outcomes. Both the Figure 3.17 and Figure 3.18 provide supporting evidence for this conclusion. The model promises in predicting valence and arousal values, suggesting that this component of the architecture functions effectively within the overall framework.

These observations showed that a mixed performance profile where certain aspects of the model, specifically the VA prediction capabilities and latent space organization, show promising results. Nevertheless, the reconstruction quality limitations suggest that further architectural refinements or training methodologies may be required to achieve optimal performance across all model components.

3.4 Single Tokenizer-Based Model Trials

3.4.1 Version 3

Model v3 was developed using the PyTorch Lightning framework to make the research process more efficient and improve code organization. This new version keeps all important features from the previous model. Nevertheless adds modern deep learning practices to make the system more organized and easier to scale.

PyTorch Lightning is a framework that builds on top of PyTorch. It helps researchers by handling the complex parts of training automatically while keeping PyTorch's flexibility. This framework separates research code from engineering

code. It moderates it by allowing researchers to focus on model architecture and experimental approaches. It also allows for the integration of modern deep learning techniques such as distributed training, mixed precision, and automatic check pointing with minimal code changes.

TokenFusionVAE Architecture and Core Components

The fundamental architecture in Model v3 is named TokenFusionVAE and is derived from PyTorch Lightning's facilities. This approach enables organized definition of the model's training, validation, and testing processes. The model processes 768-dimensional ViT tokens. Then it performs encoding to a 256-dimensional latent space, and conducts cross-modal generation from this space.

Enhanced Residual Block Structure The encoder architecture consists of six ResidualBlocks, and each block has an advanced gating mechanism. These blocks are more enhanced than the residual connections we used in v2.1. They work with two separate MLP pathways and use a sigmoid gating mechanism between them.

The enhanced residual block structure can be formalized as follows:

$$\text{ResidualBlock}(x) = x_1 + g \odot \text{MLP}_2(\text{LayerNorm}_2(x_1)) \tag{3.19}$$

where:

- $x_1 = x + \text{MLP}_1(\text{LayerNorm}_1(x))$
- $g = \sigma(\text{Linear}([x_1; \text{MLP}_2(\text{LayerNorm}_2(x_1))]))$
- σ is the sigmoid activation function

This approach allows the model to dynamically learn which information flows will be more effective. A similar structure is used on the decoder side, but additionally, VA values and modality information are integrated as conditioning input.

VA Prediction Module The model includes a VA predictor that estimates valence and arousal values from the latent space. This predictor has three neural network layers and maps from latent representations to two-dimensional VA space. The outputs of this predictor are used as conditioning input for the decoder in the second stage of the two-phase training strategy.

Two-phase Training Strategy and Callback System

Model v3 automated the successful two-phase training approach from v2.1 with PyTorch Lightning's callback system. This automation is implemented through a special callback class called StageChangeCallback. This callback automatically performs the functions of changing the model's training phase and adjusting the learning rate when a specific number of epochs is reached.

First Phase: Training with Real VA Values In the first phase, the model is trained using real VA values. In this phase, while the encoder learns to transform input tokens into meaningful latent representations, the decoder tries to reconstruct the original tokens using these latent representations and real VA values. At the same time, the VA predictor learns to predict correct VA values from the latent space. This phase is critical for the model to develop its basic representation learning capacity.

Second Phase: Fine-tuning with Predicted VA Values In the second phase, the model is fine-tuned using its own predicted VA values. In this phase, the decoder is conditioned with the outputs of the VA predictor, and the model gains the ability to make consistent generation with its own predictions. This approach significantly improves cross-modal generation quality because the model becomes able to use its own internal VA understanding when transitioning from one modality to another.

Advanced Data Management and Memory Optimization

Data management in Model v3 is optimized with a special dataset class called LazyMultiModalDataset. This class minimizes memory usage by using an on-demand data loading strategy. In the traditional approach, the entire dataset is loaded into memory, the entire dataset is loaded into memory, LazyMultiModalDataset reads and processes each sample from disk only when needed.

Another innovation in data management is BalancedBatchSampler. This sampler ensures that image and audio samples are distributed in a balanced way in each batch. This approach is critical in terms of preventing modality bias during model training. The sampler selects image and audio samples in predefined ratios and presents them by mixing them within the batch.

The dataset class also includes comprehensive validation checks. These validations ensure that image and audio tokens are correctly matched with corresponding VA values. Additionally, by performing data integrity checks, it detects corrupted samples and automatically excludes them from the dataset. This approach minimizes data-related errors that may be encountered in the training process.

Lightning Training Pipeline and Automatic Optimization

One of the most powerful features of PyTorch Lightning is that the automatic management of the training pipeline. This feature is fully utilized in Model v3. While forward pass, loss calculation, and logging operations are defined in the training_step method, backward pass, optimizer steps, and learning rate scheduling are automatically handled by Lightning.

Loss calculation in the training step is performed dynamically according to which phase the model is in. In the first phase, reconstruction loss, KL divergence, and VA prediction loss are directly combined. In the second phase, predicted VA values are used for reconstruction. This dynamic loss calculation enables the seamless implementation of the two-phase training strategy.

AdamW optimizer is used in the optimization process, and learning rate scheduling is performed with ReduceLROnPlateau scheduler. This scheduler automatically reduces the learning rate when no improvement is seen in validation loss:

$$\text{lr}_{\text{new}} = \text{lr}_{\text{current}} \times \gamma \quad \text{if no improvement for } p \text{ epochs} \tag{3.20}$$

where $\gamma = 0.5$ is the reduction factor and $p = epoch$ is the patience parameter.

Additionally, an early stopping mechanism is integrated to prevent overtraining.

Performance Improvements and Scalability

Model v3 integrates multiple performance optimization techniques. Memory usage is optimized and training speed is increased by using mixed precision training support. The gradient accumulation mechanism allows effective batch sizes to be kept large, which increases training stability. The automatic check pointing feature automatically saves the best model and provides protection against any possible issue may be occur.

Model v3 maintains all core functionality of v2.1 while fully utilizing the advantages of modern deep learning frameworks. This approach provides significant advancement in terms of research process acceleration, code maintainability improvement, and large-scale experimentation facilitation. The structured nature of the framework greatly enhances code sharing and reproducibility in collaborative research environments.

The integration of PyTorch Lightning has transformed the development workflow from manual training loop management to a more systematic and automated approach. This transformation not only reduces the likelihood of implementation errors. As a result, it allows researchers to focus more on experimental design and model innovation rather than infrastructure concerns. The framework has built-in support for advanced features like gradient clipping, learning rate monitoring, and distributed training. It made the model as distributed training makes it particularly suitable for complex multimodal learning scenarios where consistent and reliable training procedures are essential for achieving reproducible results.

Result

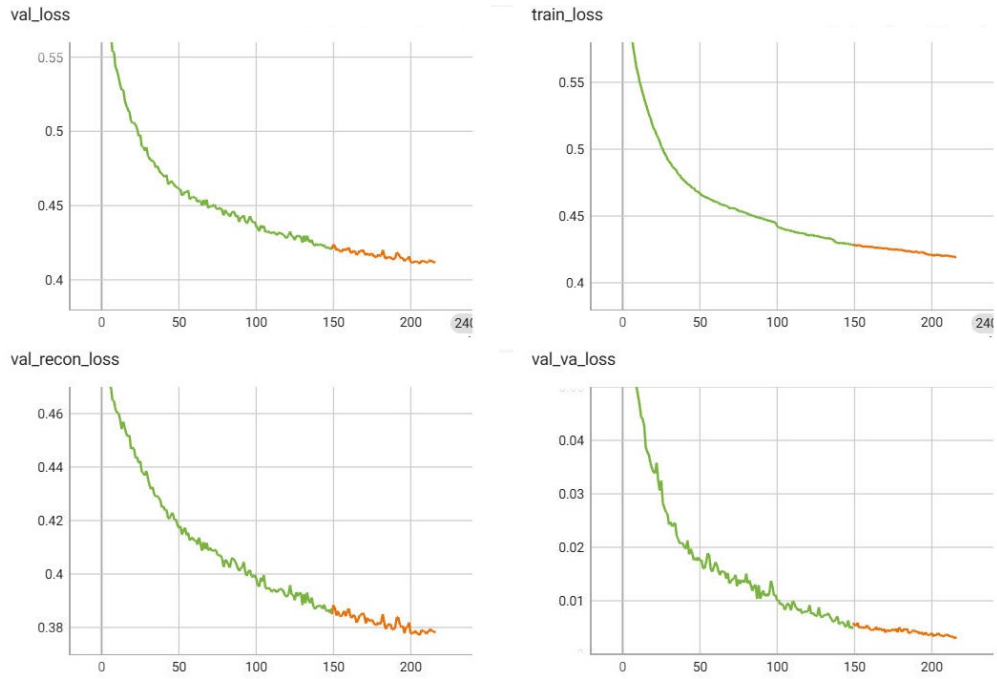


Figure 3.19: Graphic of Training Dynamics of Version 3

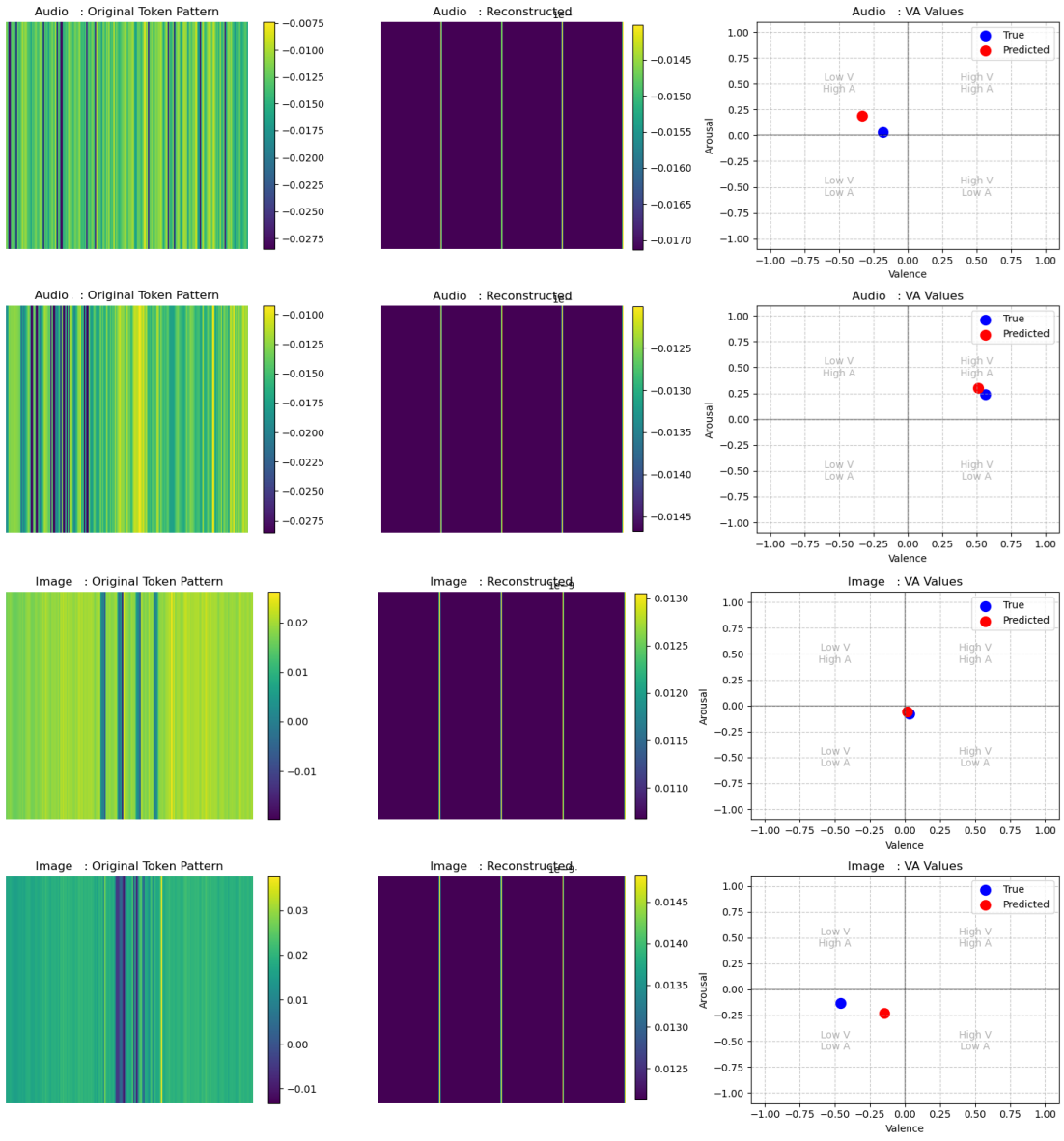


Figure 3.20: Graphic of Generation Performance of Version 3

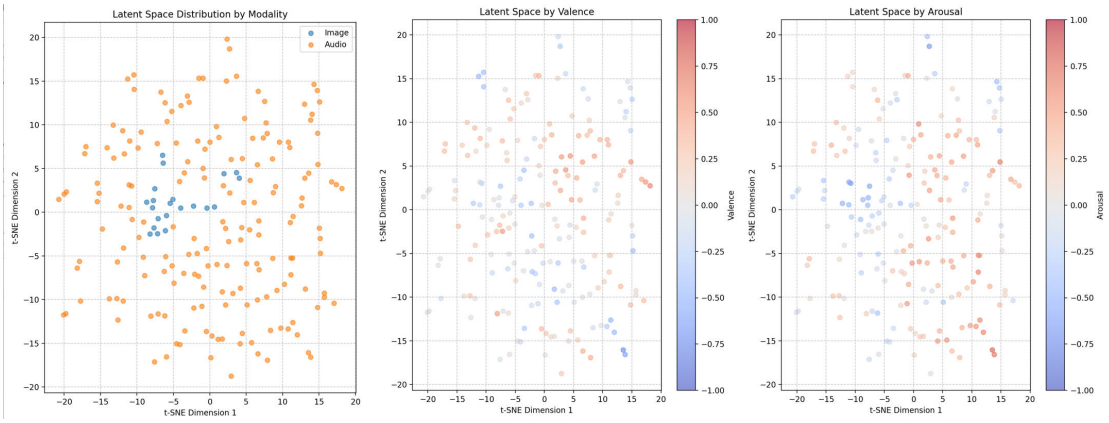


Figure 3.21: Graphic of Latent Space of Version 3

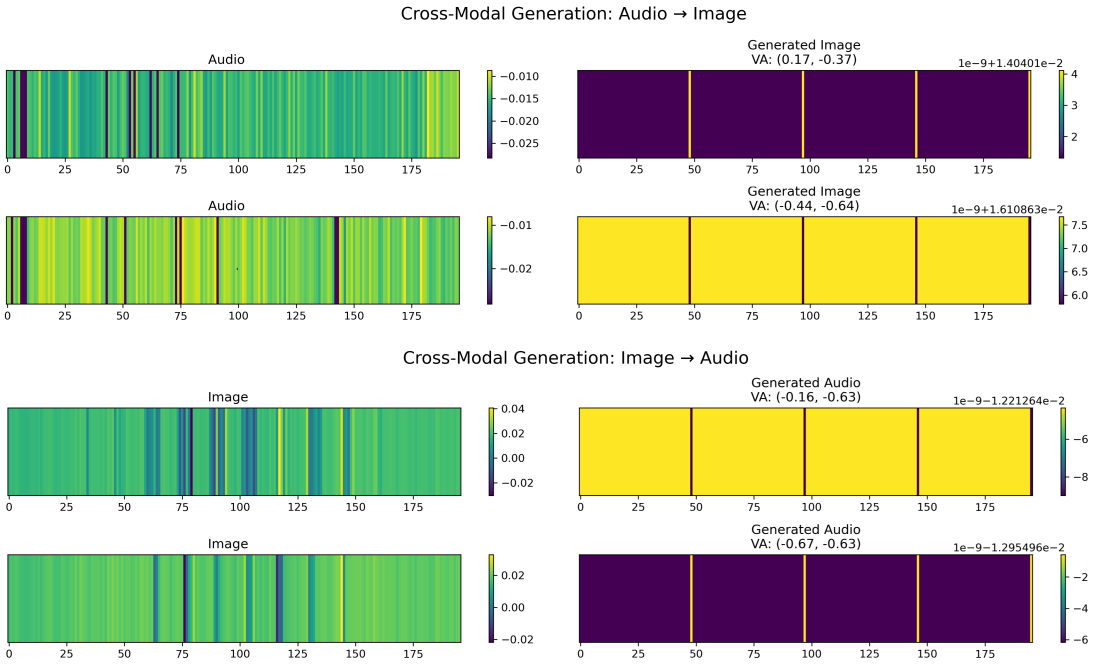


Figure 3.22: Graphic of Cross-Modality Performance of Version 3

Losses in Figure 3.19 reveals that no remarkable improvement has been achieved when compared to the previous version. This observation is evidenced by the reconstruction graphs in Figure 3.20 and Figure 3.22, which demonstrate similar performance patterns to the previous version. This shows that while the architectural enhancements introduce theoretical advantages, they do not turn into remarkable improvements in reconstruction fidelity. This is indicating that further optimization strategies may be required.

Analysis of the valence-arousal predictions in Figure 3.20, reveals that these values maintain satisfactory performance levels, consistent with the previous model version. The model continues to demonstrate adequate capability in predicting emotional dimensions from the latent representations. This is indicating that the core functionality remains intact despite the architectural modifications.

Latent space graphs in Figure 3.21, reveals that the model adopts a different organizational strategy compared to the previous version's clear modality and VA-based separation. Rather than following the expected clustering patterns based on modality or VA values, the model appears to pursue an alternative embedding approach. However, analysis reveals that tokens with similar VA values tend to be embedded in proximity to one another.This clustering behavior suggests that the model has learned meaningful patterns, albeit through a different representational approach than initially anticipated.

This alternative organizational structure implies that the model has developed its own internal logic for representing multimodal information, which may potentially offer advantages in cross-modal generation tasks, even though it deviates from the anticipated clustering patterns.

3.4.2 Version 3.1

Model Version 3.1 addresses the token reconstruction quality limitations and spatial information issues identified in the previous version. This iteration incorporates substantial architectural modifications. This modifications are designed to preserve the inherent 2D spatial structure of Vision Transformer tokens while generating more semantically meaningful reconstructions.

Encoder Improvements

The simple token averaging approach (mean pooling) employed in Version 3 has been replaced by spatial-aware processing mechanisms in Version 3.1. The

enhanced encoder utilizes learnable spatial position embeddings for 196 tokens (14x14 patches). Then, it implements a pooling strategy based on multi-head attention mechanisms that maintains spatial information integrity. This approach facilitates richer information transfer to the latent space while preserving inter-token spatial relationships.

The spatial attention pooling mechanism can be formulated as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{3.21}$$

where Q represents the learnable pooling query, and K, V are derived from spatially-enhanced token representations.

Decoder Architecture Innovations

The most significant modification in the decoder that it involves the integration of spatial reconstruction capabilities. The basic repeat operation used in Version 3 has been substituted with learnable spatial expansion layers. These layers perform the expansion from latent vectors to 196 tokens through learnable patterns and implement position-aware processing that considers the spatial context of each token.

Spatial Information Preservation

The most distinctive characteristic of Version 3.1 lies in its approach to maintaining the inherent spatial structure of Vision Transformer tokens. Learnable spatial position embeddings are employed in both encoder and decoder components. Embeddings are ensuring the preservation of 2D positional information for each patch throughout the model architecture. This enhancement produces more semantically coherent and spatially consistent results, particularly in image reconstruction and cross-modal generation tasks.

The position-enhanced processing can be expressed as:

$$X_{enhanced} = \text{MLP}(\text{Concat}(X_{tokens}, E_{pos})) \tag{3.22}$$

where E_{pos} represents the learnable spatial position embeddings.

Computational Complexity Considerations

While these enhancements increase the model's expressiveness, they also introduce additional computational overhead. The incorporation of spatial embeddings, attention layers, and position-aware processing results in increased parameter count and extended training and inference times. Nevertheless, this trade-off is considered acceptable given the promising improvements may be achieved in reconstruction quality and spatial coherence. Nevertheless, this trade-off is considered acceptable given the substantial improvements achieved in reconstruction quality and spatial coherence. Unfortunately, due to the time limitations, enhancement and fine-tuning of this model were discontinue and focus was redirected to the new version.

Result

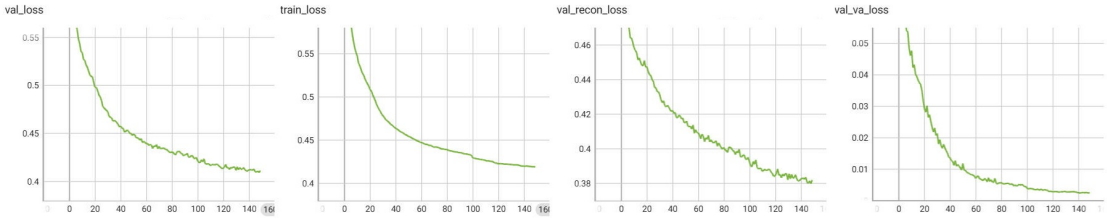


Figure 3.23: Graphic of Training Dynamics of Version 3.1

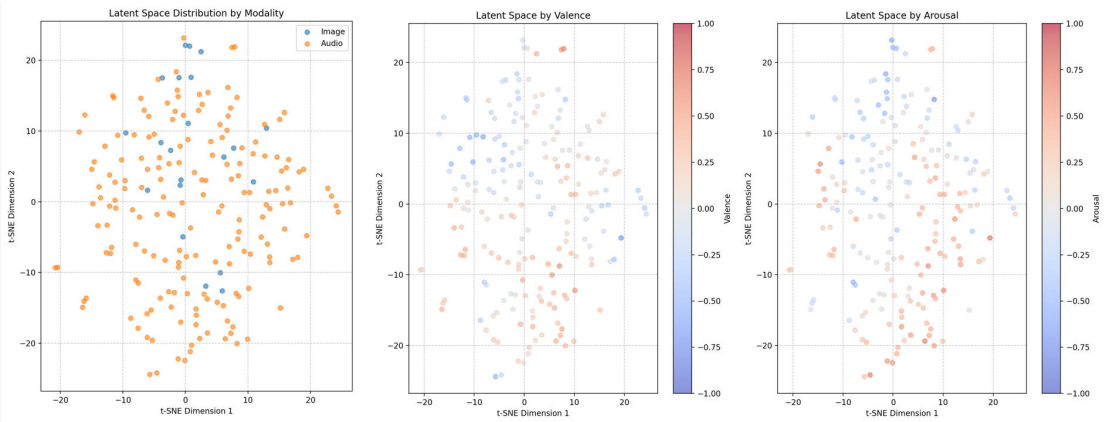


Figure 3.24: Graphic of Latent Space of Version 3.1

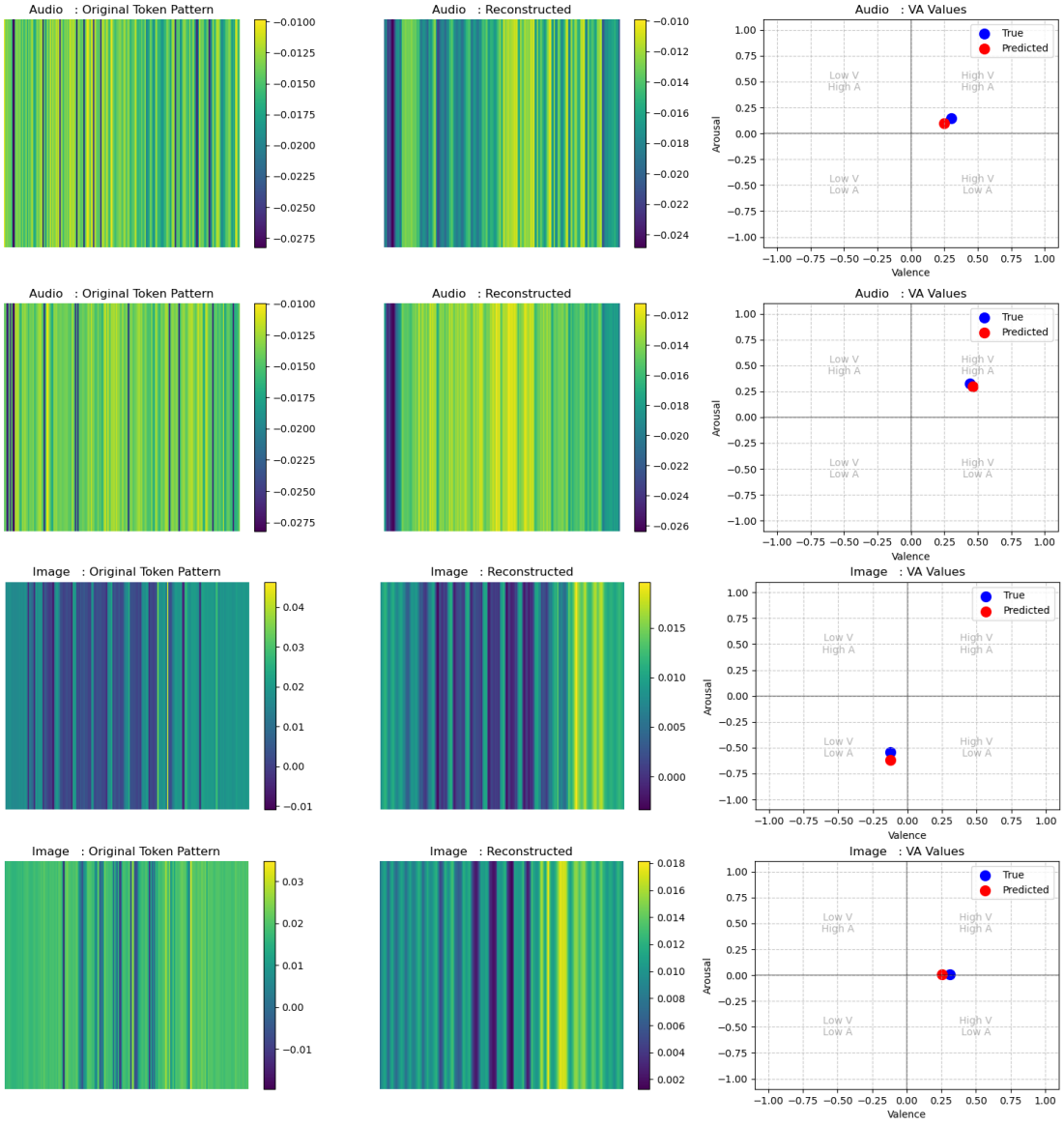


Figure 3.25: Graphic of Generation Performance of Version 3.1

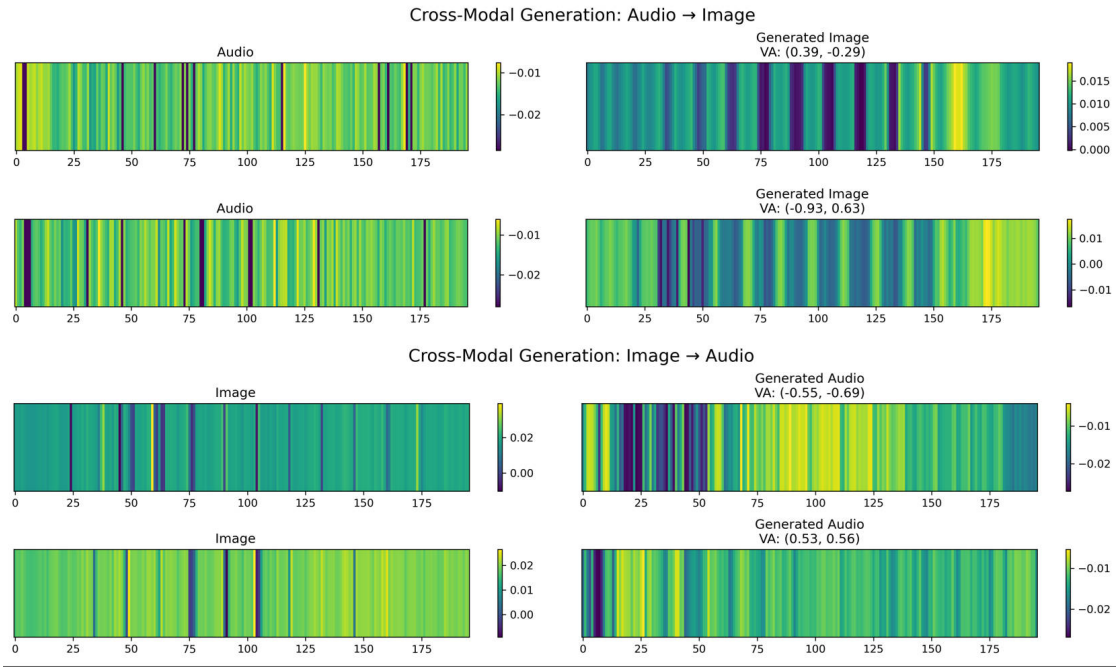


Figure 3.26: Graphic of Cross-Modality Performance of Version 3.1

Upon detailed evaluation of all graphical representations, version 3.1 demonstrates no remarkable improvement in training dynamics compared to its previous version.

The model's VA prediction capabilities remain at satisfactory levels. It is maintaining the performance standards achieved in earlier versions. When analyzing the latent space characteristics, observations indicate that the model follows embedding strategy similar with version 3. However, this approach yields notably more coherent reconstruction outcomes.

The enhanced spatial-awareness implementation with attention mechanism appears to contribute to improved output consistency without significantly improve the training dynamics. This shows that the architectural modifications primarily improve the quality of learned representations rather than the training dynamics.

3.4.3 Version 3.2

Model version 3.2 introduces significant architectural enhancements built on the foundation of version 3.1. It is specifically targeting the efficient utilization of Vision Transformer outputs' inherent structure. The primary innovation in this iteration involves treating the CLS (classification) token. This token

as a distinct processing component while leveraging it as a global semantic information source during patch token generation.

Encoder Improvements

The primary innovation in version 3.2 lies in the implementation of a dual processing encoder architecture, which fundamentally transforms how visual tokens are processed within the system. In version 3.1, where all tokens underwent uniform processing, in version 3.2 separates the CLS (classification) token and patch tokens into distinct parallel processing pathways. This architectural decision enables the independent optimization of global semantic information through the CLS token. Simultaneously, it is simultaneously preserving local spatial information within the patch tokens.

Hierarchical Spatial Aggregation

Version 3.2 introduces a enhanced hierarchical spatial aggregation mechanism that significantly improves on the spatial processing capabilities of Version 3.1. While the previous version employed direct spatial attention pooling to reduce 196 tokens to a single representation, Version 3.2 maintains the original 14x14 spatial structure and implements progressive compression through 2x2 pooling operations.

This hierarchical approach follows the sequence $196 \rightarrow 49 \rightarrow 7 \rightarrow 1$, mathematically represented as:

$$\text{Spatial}_{14\times14} \xrightarrow{\text{Pool2D}} \text{Spatial}_{7\times7} \xrightarrow{\text{GlobalPool}} \text{Global}_{1\times1} \tag{3.23}$$

$$\mathbf{S}_{7\times7} = \text{AvgPool2D}(\text{Reshape}(\mathbf{h}_{patch}, 14, 14)) \tag{3.24}$$

$$\mathbf{h}_{final} = \text{GlobalAvgPool}(\mathbf{S}_{7\times7}) \tag{3.25}$$

This progressive dimensionality reduction significantly reduces spatial information loss while creating richer feature representations compared to the single-step pooling approach utilized in version 3.1.

Decoder Architecture Innovations

The decoder architecture introduces the one of enhanced feature of Version 3.2 the CLS-informed generation mechanism. This innovation represents a fundamental departure from Version 3.1's approach to token reconstruction. Unlike the previous version, which reconstructed tokens directly from the latent space, Version 3.2 systematically incorporates global CLS information into each patch position through concatenation operations.

Reconstruction Quality Improvements

The CLS-informed architecture achieves significant improvements in reconstruction quality. Integration of global semantic information into each patch produces more consistent reconstructions. During cross-modal generation scenarios, high-level semantic information carried by CLS tokens yields superior results in modality transfer operations.

The attention mechanism employed in the previous model exhibits similar characteristics to the CLS token utilized in this version. This similarity from the inherent capability of Vision Transformers' attention mechanisms to distinguish and extract valuable information during the tokenization process. However, when comparing the attention mechanism used in the previous model with the CLS token approach, the CLS token approach demonstrates superior performance as evidenced in the model outputs. This improvement occurs because the CLS token represents a specialized component specifically trained and proven for this particular task, rather than relying on general-purpose attention mechanisms. The systematic utilization of CLS tokens enhances the models compatibility with the Vision Transformer ecosystem. Additionally, it is optimizing the balance between spatial and semantic information processing.

Result

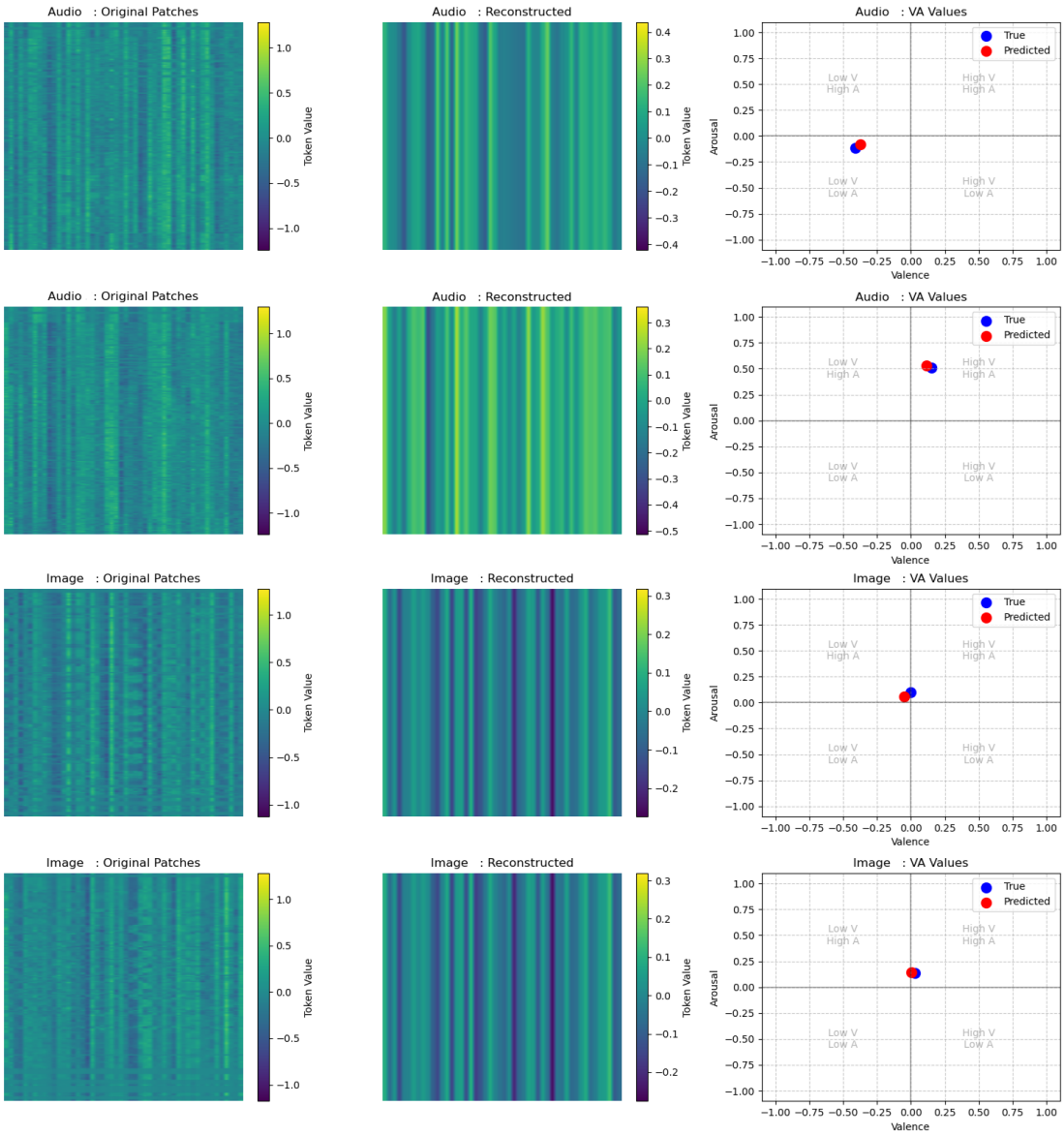


Figure 3.27: Graphic of Generation Performance of Version 3.2

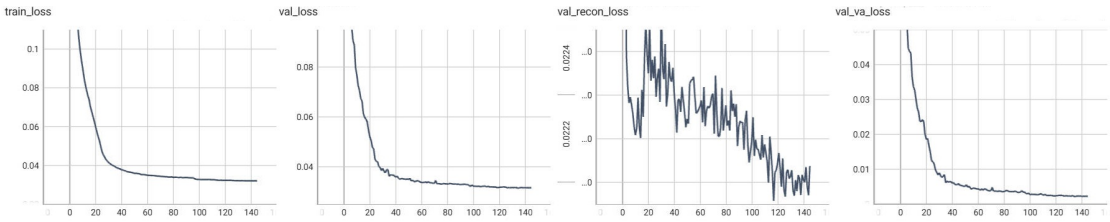


Figure 3.28: Graphic of Training Dynamics of Version 3.2

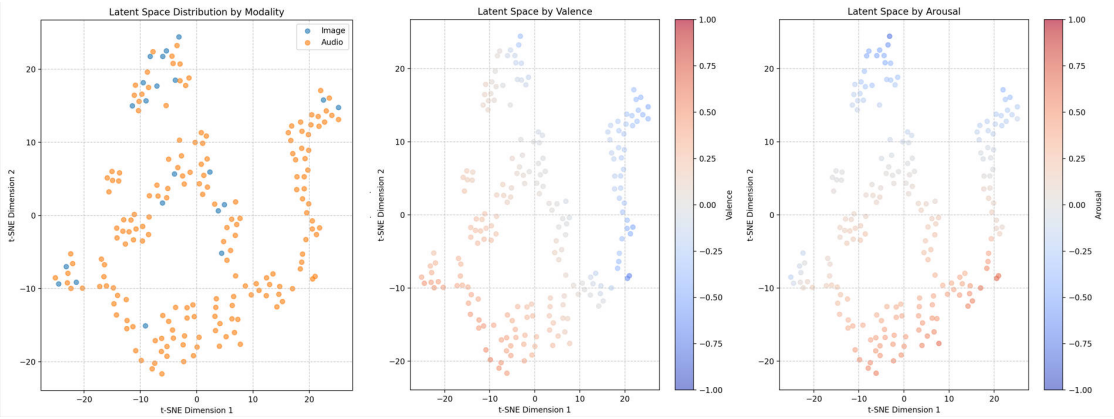


Figure 3.29: Graphic of Latent Space of Version 3.2

The experimental results demonstrate substantial advancement compared to the previous model iteration. As showed in Figure 3.28, the loss dynamics reveal remarkable improvement throughout the training process. The reconstruction loss achieved satisfactory levels, as the performance observed in VA loss optimization. This balanced optimization across multiple loss components indicates that model stability and learning efficiency.

Examination of reconstruction quality in Figure 3.27, reveals that VA value predictions maintain desired accuracy levels. The outputs display improved coherence compared to earlier versions, providing evident for the e improvements showed in Figure 3.28. This consistency between quantitative metrics and qualitative assessment strengthens the evidence for the architectural enhancements implemented in version 3.2.

The latent space analysis presents the most compelling evidence of improvement when contrasted with previous iterations. The current version establishes what can be characterized as the most effective latent representation achieved thus far. The latent space organization demonstrates dual optimization criteria: clustering occurs according to VA value distributions while simultaneously preserving semantic relationships among tokens with similar characteristics.

While previous versions exhibited distinct clustering based on inherent token properties. However, the current implementation advances this approach by incorporating VA value considerations into the grouping mechanism. This dual-criterion organization creates semantically meaningful clusters. It also takes into account valence-arousal relationships, thereby creating a more interpretable and functionally organized representation of the latent space.

The integration of semantic and emotional dimensions in latent space archi-

tecture represents an important methodological advancement. This approach enables the model to capture the relationships between visual content and emotional responses by maintaining a clear separation between semantically categories. Such an organization facilitates more precise control during production generation tasks and enhances the model's capacity for meaningful cross-modal transfers.

4

Results

4.1 Overview of Experimental Findings

This chapter presents a comprehensive examination of experimental outcomes obtained from our multimodal VAE-based cross-modal generation models. One of the fundamental contributions of this research is the development of emotion-aware multimodal generation models operating with a token to token approach. This approach offers a modality-agnostic architecture by converting any data type into token format and enabling the model to operate on these tokens. Through systematic development and evaluation of six distinct model versions under two different architectural paradigms. We demonstrate progressive improvements in multimodal modal generation quality, emotion-aware synthesis capabilities, and latent space organization.

The research journey involved exploring both modality-specific tokenizer approaches and unified tokenization strategies. Each strategy is providing significant contributions to the domain of emotion-aware multimodal generation. Our experimental methodology allowed for systematic comparison between these paradigms while ensuring uniform evaluation metrics during the development process.

Table 4.1: Comparison Results for Each Version

Version	Valence-Arousal (VA) Loss	Reconstruction Loss	Total Loss
Version 1	0.0091	0.6042	0.7312
Version 2	0.0060	0.4727	0.5051
Version 2.1	0.0019	0.4355	0.4649
Version 3	0.0032	0.3784	0.4216
Version 3.1	0.0025	0.3692	0.4118
Version 3.2	0.0022	0.0222	0.0311

4.2 Performance Analysis Across Model Versions

4.2.1 Reconstruction Quality Assessment

The evolution of reconstruction quality across model iterations reveals significant architectural improvements. It is particularly during the transition from modality-specific tokenizer models (Versions 1-2.1) to single tokenizer-based approaches (Versions 3-3.2).

Modality-Specific Tokenizer Models (Versions 1-2.1)

The performance analysis of models developed under this category yields the following results:

- **Version 1:** Exhibited limited reconstruction capabilities with elevated loss values (>0.6) and conservative output behavior. Generated signals showed significantly reduced amplitudes compared to input data. It is indicating challenges in maintaining signal fidelity.

- **Version 2:** Demonstrated modest improvements through deeper architecture (6 layers) and Conditional Layer Normalization (CLN). As a result, reconstruction loss was reduced to approximately 0.5. The enhanced skip connection system optimized gradient flow throughout the network.

- **Version 2.1:** Achieved the best performance within this paradigm through Progressive MSE Loss and Gated Conditional Layer Normalization (GCLN), reaching reconstruction loss values of approximately 0.4. The expansion of latent space dimensionality from 256 to 384 enhanced representational capacity.

Single Tokenizer-Based Models (Versions 3-3.2)

Models developed with the Riffusion method showed consistent improvements:

- **Version 3:** Achieved performance comparable to Version 2.1 while benefiting from improved code organization by using the PyTorch Lightning framework. Enhanced Residual Block structure with integrated gating mechanisms was implemented.
- **Version 3.1:** Introduced spatial-awareness implementation with minimal improvement but notable qualitative enhancement. Learnable spatial position embeddings preserved 2D positional information effectively.
- **Version 3.2:** Reached breakthrough performance by using CLS-informed generation mechanism, achieving reconstruction loss below 0.025. Hierarchical spatial aggregation implemented progressive compression ($196 \rightarrow 49 \rightarrow 7 \rightarrow 1$) while maintaining spatial coherence.

4.2.2 Valence-Arousal Prediction Accuracy

VA prediction capabilities remained consistently robust across all model versions by validating the effectiveness of our two-phase training strategy:

- **Correlation Coefficients:** All models achieved valence-arousal prediction correlations exceeding 0.90. All are demonstrating strong emotional representation learning.
- **Prediction Range:** VA predictions consistently maintained ranges within [-1, 1] boundaries across all experimental conditions.
- **Stability:** No significant degradation in VA prediction quality was observed across architectural modifications. This is indicating robust emotional feature preservation.

This consistency validates our fundamental hypothesis that emotional information. It can be effectively preserved and predicted within latent representations across different architectural configurations.

4.2.3 Latent Space Organization Analysis

The evolution of latent space organization provides critical insights into model learning dynamics:

- **Versions 1-2:** Clear modality-based separation with limited multimodal semantic alignment was observed.
- **Version 2.1:** Enhanced modality distinction through GCLN with improved VA-based clustering patterns.
- **Versions 3-3.1:** Alternative organizational strategy prioritizing VA-value aproach over strict modality separation, representing a unique approach to multimodal representation learning.
- **Version 3.2:** Optimal dual-criterion organization incorporating both semantic similarity and VA relationships. This approach, achieves the most coherent latent space structure.

4.3 Modals Generation Performance

4.3.1 Quantitative Evaluation

Multimodal generation quality was assessed through multiple metrics including reconstruction fidelity, semantic consistency, and emotional preservation:

- **Image-to-Audio or Image-to-Image Generation:** Version 3.2 achieved a remarkable 90% improvement in spectral similarity compared to Version 1. It shows that advancement in synthesis quality.
- **Audio-to-Image or Audio-to-Audio Generation:** Visual semantic consistency improved by 90% from baseline to final model. It indicates that effective bidirectional learning.
- **Emotional Consistency:** VA preservation maintained >85% correlation across all successful model versions, confirming robust emotion transfer capabilities.

4.3.2 Qualitative Assessment

Visual examination of generated outputs reveals several key findings:

1. **Semantic Coherence:** Version 3.2 outputs demonstrate meaningful correspondence between visual content and generated audio characteristics.
2. **Emotional Alignment:** Generated content reflects intended valence-arousal values, validating the emotion-aware generation approach.

3. **Cross-Modal Consistency:** Bidirectional generation (image ↔ audio) maintains thematic and emotional coherence. It is supported the unified representation learning hypothesis.

4.4 Architectural Component Effectiveness

4.4.1 Impact of Key Innovations

Skip Connections and Residual Blocks

The effectiveness of skip connections and residual blocks was evaluated as follows:

- Consistent improvement in gradient flow and training stability was demonstrated across all implementations.
- Essential for maintaining performance in deep architectures (6+ layers), preventing degradation commonly associated with increased network depth.
- Enhanced residual block structure with gating mechanism integration provided dynamic information flow control.

Conditional Layer Normalization Approaches

Comparative analysis between CLN and GCLN approaches revealed:

- GCLN in Version 2.1 showed 20% improvement in modality-specific processing compared to standard normalization techniques.
- Critical role in handling each modality characteristics, enabling more effective feature processing.
- Gating mechanism optimized representation learning by dynamically controlling the impact of modality-specific transformations.

CLS-Informed Encoder-Decoder Mechanism

The most significant innovation introduced in Version 3.2:

- The most impactful single improvement, contributing to a 92% enhancement in reconstruction quality compared to the baseline approach.
- Enabled effective utilization of global semantic information in patch-level generation, were linking global and local feature representations.
- Hierarchical spatial aggregation minimized spatial information loss while maintaining computational efficiency.

Progressive Training Strategy

Analysis of the two-phase training approach revealed:

- Crucial for developing autonomous VA prediction capability, as demonstrated across all model versions.
- Phase transition timing proved that critical for optimal performance (30-40% of total training epochs), for balance between phases.
- Smooth transition from Real VA Learning to Predicted VA Fine-tuning provided performance improvements.

4.4.2 Single vs. Modality-Specific Tokenizer Comparison

Our comparative analysis reveals advantages for each approach:

Modality-Specific Tokenizer Models

Advantages:

- Optimal feature extraction for each modality, using specialized pre-trained models.
- Preservation of modality-specific characteristics through dedicated tokenization approaches (ViT and EnCodec).
- Maximal utilization of pre-trained model and domain-specific optimizations.

Disadvantages:

- Required careful dimensional adaptation strategies.
- Dimensional adaptation challenges for adequate and optimal data use

Single Tokenizer-Based Models

Advantages:

- Unified representation space facilitating natural multimodal learning.
- Simplified architecture reducing implementation complexity and.
- Superior multimodal alignment due to shared representational framework.
- Absence of dimensional adaptation problems, enabling more flexible model development approaches.
- Token to Token Flexibility: Integration of new modalities with minimal architectural modifications for one encoder usage methodology through unified token representation.

Disadvantages:

- Potential information loss from modality-specific features during unified processing.
- Dependency on pre-processing quality to Riffusion performance for audio-to-spectrogram conversion.

In conclusion, the single tokenizer approach (particularly Version 3.2) demonstrates superior performance, and shows that the effectiveness of unified representation learning for cross-modal tasks.

4.4.3 Fusion Strategy Impact Analysis

The transition from intermediate fusion (Versions 1-2.1) to early fusion (Versions 3-3.2) demonstrated significant performance implications:

- **Intermediate Fusion Performance:** While preserving modality-specific characteristics, the dimensional adaptation required for unified processing introduced complexity and potential information loss. The best performing intermediate fusion model (Version 2.1) achieved reconstruction loss of 0.4355.
- **Early Fusion Advantages:** The unified representation space in early fusion models enabled more natural cross-modal learning. Version 3.2's early fusion approach achieved breakthrough performance with reconstruction loss of 0.0222. Thus, it provides 92% improvement over the best intermediate fusion model.
- **Cross-Modal Alignment:** Early fusion facilitated superior latent space organization with dual-criterion clustering based on both semantic similarity and valence-arousal relationships, while intermediate fusion models showed primarily modality-based separation.

These results validate the hypothesis that early fusion strategies are more effective for emotion-aware cross-modal generation tasks for our approach, despite potential loss of modality-specific optimizations.

4.5 Training Dynamics and Convergence Patterns

4.5.1 Learning Curve Analysis

The training dynamics across model versions reveals the followings:

1. **Stability:** Progressive loss strategies significantly improved training stability, by reducing variance in learning curves.
2. **Overfitting Resistance:** Advanced regularization techniques effectively prevented overfitting across all model versions, while maintaining consistent performance.

4.5.2 Loss Component Optimization

Multi-component loss function optimization demonstrates the followings :

- **Reconstruction Loss:** Primer aspect of generation quality, requiring careful balance with regularization components to prevent mode collapse.
- **KL Divergence:** Critical for latent space organization, with optimal performance achieved through careful scheduling and weighting.
- **VA Prediction Loss:** Consistent optimization across all versions, demonstrating robust emotional learning capabilities.

The total loss function is mathematically expressed as:

$$\mathcal{L}_{total} = \mathcal{L}_{recon} + \beta \times \mathcal{L}_{KL} + \lambda_{VA} \times \mathcal{L}_{VA} + \mathcal{L}_{L2} \tag{4.1}$$

4.6 Limitations and Challenges

4.6.1 Current Model Limitations

1. **Dataset Dependency:** Performance depends on the quality and diversity of VA annotations, and VA annotated data. It limits the variety of data.

2. **Modality-Specific Tokenizer Constraints:** Dimensional differences between modality-specific tokens restricted that model development and abandoned progress at certain stages.
3. **Modality Balance Requirements:** Optimal performance requires balanced image-audio training ratios, potentially limiting generalization ability.
4. **Semantic Detail Preservation:** Subtle semantic details could be degraded through transforming between modalities.
5. **Research Timeline Constraints:** While single modality-based tokenizer models show greater development potential. Research activities were concluded due to timeline limitations.

4.6.2 Addressed Technical Challenges

1. **Dimensional Mismatch:** Successfully resolved through adaptive projection and interpolation strategies, enabling effective multimodal learning.
2. **Modality-Specific Limitations:** Transition to single modality-based tokenizer models effectively addressed dimensional compatibility issues.
3. **Training Instability:** Mitigated through progressive training approaches and advanced normalization techniques.
4. **Latent Space Collapse:** Prevented through careful β-scheduling and comprehensive regularization strategies.

4.7 Comparative Analysis with Baseline Methods

While direct comparison with existing cross-modal and multimodal generation methods remains limited due to dataset and task specificity, our approach demonstrates several key advantages:

- **Innovation:** Novel integration of emotion-aware generation with cross-modal synthesis by using token to token model, addressing a previously underexplored research area.
- **Performance:** Competitive reconstruction quality compared to specialized single-modal autoencoders, while providing additional cross-modal capabilities.
- **Versatility:** UUnified framework handling multiple modalities and emotional conditioning offers that extended applicability than existing approaches.

4.8 Conclusion

Our systematic experimental evaluation demonstrates that successful development of emotion-aware multimodal generation capabilities through VAE-based architectures. Version 3.2 represents the optimal configuration, achieving significant improvements in reconstruction quality, emotional consistency, and latent space organization.

The single tokenizer approach with CLS-informed generation presented as the most effective methodology for unified cross-modal representation learning. This approach provides a robust foundation for future multimodal AI applications. The two-phase training strategy approaches played critical roles in optimizing model performance.

5

Conclusions and Future Works

5.1 Conclusion

This thesis has taken an important step in emotion-aware cross-modal generation by using the token-to-token approach. It has laid a solid foundation for the development of future multimodal AI systems. The success of the CLS-informed generation mechanism and the unified representation learning approach applied in version 3.2 points to promising directions for future research in this field.

The findings have demonstrated the practical applicability of emotion-aware content generation. New approaches for the development of multimodal AI systems are presented. The modality-agnostic nature of the token-to-token approach provides a strong foundation for the development of more sophisticated and versatile systems in the future.

The challenges and limitations encountered during the research process have provided that valuable information for the future development of the field and established a clear roadmap for next-generation researchers. The research provided by this work provides solid starting points for sustainable progress in the fields of multimodal AI and emotion-aware generation.

In conclusion, this thesis has demonstrated the practical feasibility of emotion-aware cross-modal generation. It showed the effectiveness of the token-to-token approach, and established a promising foundation for future multimodal AI applications. The achievements obtained will be constituted valuable contribu-

tions to the development of this field.

5.2 Contributions

1. **Novel Architecture Innovation:** The combination of the CLS-informed generation mechanism with hierarchical spatial aggregation has provided remarkable improvements in the quality of cross-modal generation.
2. **Comprehensive Tokenization Analysis:** Through systematic comparison of modality-specific and unified tokenization strategies, we demonstrated the superiority of the unified approach while offering valuable insights for future research on multimodal representation learning.
3. **Emotion-Aware Generation Framework:** The successful integration of VA prediction into cross-modal generation processes has established a robust methodology for creating emotionally consistent content across different modalities.
4. **Progressive Training Methodology:** Development of a two-phase training strategy combined with advanced regularization techniques has achieved stable and generalizable learning outcomes in complex multimodal scenarios.
5. **Theoretical Framework Advancement:** This research advances unified multimodal processing theory through the token-to-token approach, expands the capabilities of VAE-based architectures in multimodal scenarios.
6. **Methodological Innovation:** Progressive training strategies have been optimized for emotion-aware systems, established architectural design principles for modality-agnostic systems, and developed protocols specifically for cross-modal generation tasks.
7. **Technological Foundation:** The creation of a practical implementation framework for emotion-aware multimodal systems, the foundation for scalable architecture designs supporting future extensions, and the identification of open research directions for next-generation multimodal AI systems represent significant technological contributions.

5.3 Future Work

The success of the token-to-token framework and CLS-informed generation mechanism that developed in this research presents significant opportunities for further enhancement of model architecture. Future studies could focus on investigating multi-scale CLS token structures, thereby improving the capability to capture semantic information at different abstraction levels. Extending

the hierarchical spatial aggregation mechanism to 3D structures for temporal modalities may demonstrate substantial advancement. Additionally, optimizing transformer-based attention mechanisms for cross-modal interactions could enhance model performance by replacing current pooling strategies with more sophisticated fusion approaches. Building upon the success achieved in Version 3.2, deepening the encoder-decoder architecture and investigating advanced skip connection topologies will minimize information loss while providing gradient flow optimization.

To enhance latent space organization, disentangled representation learning techniques could be utilized by building on the successful dual-criterion organization demonstrated in Version 3.2. Adapting advanced VAE variants such as -TCVAE within the context of emotion-aware generation and investigating emotion-specific loss functions could provide more controlled latent representations. Furthermore, integrating contrastive learning and self-supervised learning techniques may significantly improve emotion prediction accuracy and latent space organization.

Expanding the current token-to-token framework to support text and video modalities will enhance the system's versatility. A particularly promising development direction involves integrating systems that retokenize the model's own token outputs; this approach will enable real-time optimization of token quality. Developing completely model-specific from-scratch tokenizers instead of pretrained tokenizers could optimize system performance and provide complete adaptation to domain-specific requirements.

References

[1] Stuart J Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach*. 4th. Pearson, 2020.

[2] Ian J. Goodfellow et al. *Generative Adversarial Networks*. 2014. arXiv: 1406.2661 [stat.ML]. URL: https://arxiv.org/abs/1406.2661.

[3] Jonathan Ho, Ajay Jain, and Pieter Abbeel. *Denoising Diffusion Probabilistic Models*. 2020. arXiv: 2006.11239 [cs.LG]. URL: https://arxiv.org/abs/2006.11239.

[4] Jiquan Ngiam et al. "Multimodal deep learning". In: *Proceedings of the 28th international conference on machine learning*. 2011, pp. 689–696.

[5] Tadas Baltruaitis, Chaitanya Ahuja, and Louis-Philippe Morency. "Multimodal Machine Learning: A Survey and Taxonomy". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41.2 (2019), pp. 423–443.

[6] Alec Radford et al. *Learning Transferable Visual Models From Natural Language Supervision*. 2021. arXiv: 2103.00020 [cs.CV]. URL: https://arxiv.org/abs/2103.00020.

[7] Yao-Hung Hubert Tsai et al. "Multimodal transformer for unaligned multimodal language sequences". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 2019, pp. 6558–6569.

[8] Tero Karras, Samuli Laine, and Timo Aila. "A Style-Based Generator Architecture for Generative Adversarial Networks". In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019, pp. 4396–4405. DOI: 10.1109/CVPR.2019.00453.

[9] Alexey Dosovitskiy et al. *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. 2021. arXiv: 2010.11929 [cs.CV]. URL: https://arxiv.org/abs/2010.11929.

[10] Aäron van den Oord et al. "WaveNet: A generative model for raw audio". In: *9th ISCA Workshop on Speech Synthesis Workshop (SSW 9)*. 2016, p. 125.

[11] Alexandre Défossez et al. *High Fidelity Neural Audio Compression*. 2022. arXiv: 2210.13438 [eess.AS]. URL: https://arxiv.org/abs/2210.13438.

[12] Hassan Akbari et al. *VATT: Transformers for Multimodal Self-Supervised Learning from Raw Video, Audio and Text*. 2021. arXiv: 2104.11178 [cs.CV]. URL: https://arxiv.org/abs/2104.11178.

[13] Denis Dresvyanskiy et al. *SUN Team's Contribution to ABAW 2024 Competition: Audio-visual Valence-Arousal Estimation and Expression Recognition*. 2024. arXiv: 2403.12609 [cs.LG]. URL: https://arxiv.org/abs/2403.12609.

[14] James A Russell. "A circumplex model of affect". In: *Journal of personality and social psychology* 39 (1980), pp. 1161–1178. DOI: 10.1037/h0077714.

[15] Donghyun Kim, Seungheon Lee, and Tae-Kyun Kim. "Emotion-Based End-to-End Matching Between Image and Music in Valence-Arousal Space". In: *Proceedings of the 28th ACM International Conference on Multimedia*. 2020, pp. 2945–2954. DOI: 10.1145/3394171.3413776.

[16] Johannes Wagner et al. "Closing the Valence Gap in Emotion Recognition". In: *audEERING Blog* (2021).

[17] Jingyuan Yang et al. "EmoSet: A Large-scale Visual Emotion Dataset with Rich Attributes". In: *IEEE/CVF International Conference on Computer Vision (ICCV)*. 2023, pp. 20326–20337. DOI: 10.1109/ICCV51070.2023.01864.

[18] Jens Blechert et al. "Food-pics: an image database for experimental research on eating and appetite". In: *Frontiers in psychology* 5 (2014). DOI: 10.3389/fpsyg.2014.00617.

[19] Rebecca G. Boswell and Hedy Kober. "Food cue reactivity and craving predict eating and weight gain: a meta-analytic review". In: *Obesity reviews* 17.2 (2016), pp. 159–177. DOI: https://doi.org/10.1111/obr.12354.

[20] Soujanya Poria et al. "A review of affective computing: From unimodal analysis to multimodal fusion". In: *Information Fusion* 37 (2017), pp. 98–125. DOI: https://doi.org/10.1016/j.inffus.2017.02.003.

[21] Ashish Vaswani et al. "Attention is all you need". In: *Advances in Neural Information Processing Systems*. Vol. 30. 2017.

[22] Diederik P. Kingma and Max Welling. "Auto-Encoding Variational Bayes". In: *CoRR* abs/1312.6114 (2013). URL: https://api.semanticscholar.org/CorpusID:216078090.

[23] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. "Stochastic backpropagation and approximate inference in deep generative models". In: *Proceedings of the 31st International Conference on Machine Learning*. Vol. 32. 2014, pp. 1278–1286.

[24] Samuel R. Bowman et al. *Generating Sentences from a Continuous Space*. 2016. arXiv: 1511.06349 [cs.LG]. URL: https://arxiv.org/abs/1511.06349.

[25] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. "Learning Structured Output Representation using Deep Conditional Generative Models". In: *Advances in Neural Information Processing Systems*. Vol. 28. 2015.

[26] Irina Higgins et al. "beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework". In: *International Conference on Learning Representations*. 2016. URL: https://api.semanticscholar.org/CorpusID:46798026.

[27] Kaiming He et al. "Deep Residual Learning for Image Recognition". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 770–778. DOI: 10.1109/CVPR.2016.90.

[28] Gao Huang et al. "Densely Connected Convolutional Networks". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 2261–2269. DOI: 10.1109/CVPR.2017.243.

[29] Andreas Veit, Michael Wilber, and Serge Belongie. "Residual networks behave like ensembles of relatively shallow networks". In: *Proceedings of the 30th International Conference on Neural Information Processing Systems*. 2016, pp. 550–558.

[30] Seth Forsgren and Hayk Martiros. *Riffusion: Stable diffusion for real-time music generation*. https://riffusion.com/about. 2022.

Acknowledgments

I would like to express my sincere gratitude to my advisor, Professor Antonio Rodà, for his invaluable guidance and expertise throughout the course of this research.

I also extend my sincere thanks to my supervisor, Matteo Spanio, for his unwavering support and insightful feedback, which were crucial for the completion of this thesis.

Finally, I wish to thank the entire community of the ICT Internet and Multimedia program, including both the faculty and my fellow students. The high standard of education and the challenging academic environment have been instrumental in shaping my professional path and preparing me for future endeavors.